13th International Seminar of Speech Production

13 – 17 may 2024 Autrans FR

BOOK OF ABSTRACTS

https://issp24.sciencesconf.org

Foreword

In our opinion among the few regular conferences devoted to research on speech production ISSP is unique, because of its format. Conceived at its origins, with a reduced number of participants, as a kind of brainstorming to think about the development of international collaborations on speech production studies, it has evolved toward a regular conference with a significantly larger number of participants. Nevertheless, it has kept the spirit of its original issue and has remained over the years a privileged place for deep constructive scientific exchanges, and, because it is also a fundamental basis for fruitful collaborations, for the development of human relations among scientists. This is why, every time where it was possible, ISSP has been organized in a unique place for scientific sessions, meals and accommodation.

ISSP2024 is in line with this idea and is organized in L'Escandille, a place where it was already organized in 1996. At this time, the number of participants was about hundred. This year, more than 270 people have registered, with more than 230 people physically present on site. Thanks to the collaboration of all the participants, we were able to accommodate almost everyone in l'Escandille.

We have received 230 abstracts that were evaluated by two reviewers, whom we thank for their valuable work. In the spirit of ISSP, we accepted abstracts as long as they presented results or prospective work likely to contribute to rich discussions on the conference topics.

The organization of this conference is the fruit of the collaboration between several of the French major laboratories in the field of speech production research: Gipsa-lab in Grenoble, the Laboratoire de Phonétique et de Phonologie and the Laboratoire de Linguistique Formelle in Paris, the Laboratoire Dynamique du Langage in Lyon, the Laboratoire Parole et Langage in Aix-en-Provence, LiLPa in Strasbourg, PRAXILING in Montpellier, LORIA in Nancy. The members of these laboratories involved in the project formed both the Organizing Committee and the Scientific Committee of the conference. We would like to thank them very much for their remarkable investment on the project. It was a marvelous journey to work with all of them.

Thanks to them, we have been able to obtain financial support from a number of institutions, which are mentioned below in the list of sponsors. This has enabled us to keep registration fees relatively low, especially for students.

With the time passing, the impression left by a conference is often largely influenced by the keynotes that were presented. We want to express our grateful thanks our six keynote speakers: in the sequential order of their presentations, **Caroline Niziolek** from University of Wisconsin—Madison, USA, **Doris Mücke** from Cologne University, Germany, **Sophie Scott** from University College London, UK, **Adrien Meguerditchian,** from Université Aix-Marseille, France, **Florencia Assaneo**, from Universidad Nacional Autónoma de México, Mexico, **Jason Shaw**, from Yale University, USA.

Obviously, the success of a conference is primarily depending on the implication of all the participants. It is now in your hands.

Thank you.

Cécile Fougeron & Pascal Perrier

Chairs of ISSP2024

Table of Contents

Organizing Committee

Anne Hermes	Laboratoire de Phonétique et Phonologie - CNRS/Université Sorbonne Nouvelle
Beatrice Vaxelaire	LiLPa - Université de Strasbourg
Cecile Fougeron	Laboratoire de Phonétique et Phonologie - CNRS/Université Sorbonne Nouvelle
Fabrice Hirsch	Praxiling - CNRS / Université Paul Valéry Montpellier 3
Fanny Guitard-Ivent	Praxiling - CNRS / Université Paul Valéry Montpellier 3
Jalal Al-Tamimi	Laboratoire de Linguistique Formelle - CNRS/Université Paris Cité,
Leonardo Lancia	Laboratoire Parole et Langage - CNRS/Aix-Marseille Université
Maeva Garnier	Gipsa-lab - INP/CNRS/Université Grenoble Alpes
Mélanie Canault	Institut des Sciences et Techniques de la Réadaptation Université Claude Bernard, Lyon 1Laboratoire Dynamique du Langage - CNRS/Université Lumière, Lyon 2
Pascal Perrier	Gipsa-lab - INP/CNRS/Université Grenoble Alpes
Pierre Baraduc	Gipsa-lab - INP/CNRS/Université Grenoble Alpes
Rudolph Sock	LiLPa - Université de Strasbourg
Slim Ouni	LORIA - CNRS/Inria/Université de Lorraine
Takayuki Ito	Gipsa-lab - INP/CNRS/Université Grenoble Alpes
Véronique Boulenger	Laboratoire Dynamique du Langage - CNRS/Université Lyon 2
Yohann Meynadier	Laboratoire Parole et Langage - CNRS/Aix-Marseille Université
Yves Laprie	LORIA - CNRS/Inria/Université de Lorraine

Sponsors


CARSTENS


LABORATOIRE DE PHONETIQUE ET PHONOLOGIE
LPP


gipsa-lab


PRAXILING
Laboratoire de linguistique.


lilpa
linguistique, langues, parole


Laboratoire
Parole et Langage
50
1972
2022


Laboratoire de linguistique formelle


DDL
LABORATOIRE
DYNAMIQUE
DU LANGAGE


UNIVERSITÉ
LUMIÈRE
LYON 2


LABEX
ASLAN
UNIVERSITÉ DE LYON


Sorbonne
Nouvelle
université des cultures


Université
de Strasbourg


UGA
Université
Grenoble Alpes


LABEX
ASLAN
UNIVERSITÉ DE LYON

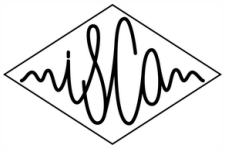

Sociétés et Humanités
Université Paris Cité


UNIVERSITÉ
PAUL
VALÉRY
MONTPELLIER 3


CerCoG
Université Grenoble Alpes


cnrs


GRENOBLE
INP
UGA

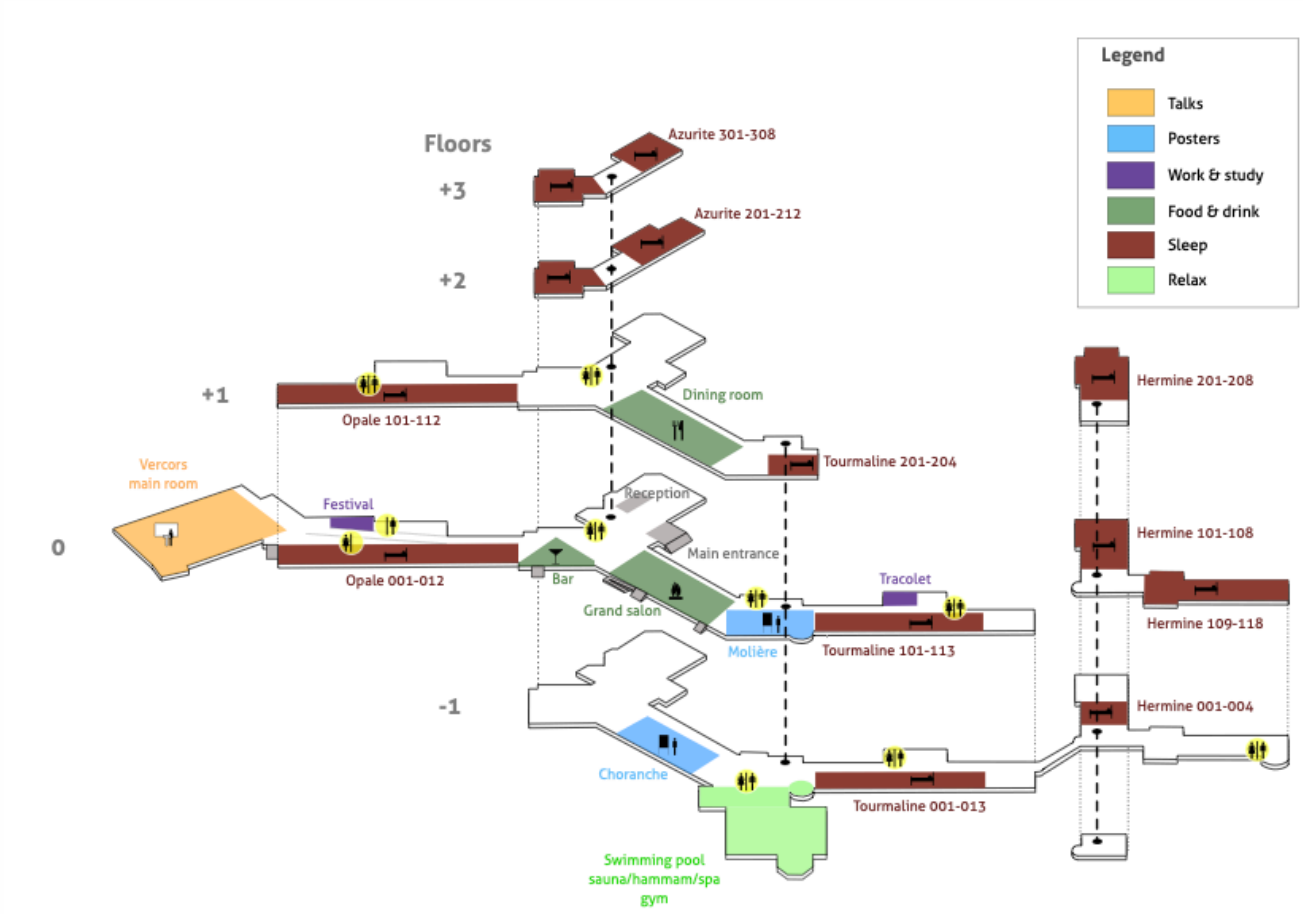



Labex
EFL


AFCP


Loria
Laboratoire lorrain de recherche
en informatique et ses applications

General information



Social programme

Monday, May 13, 2024

6:30 pm

Registration opens

8:00 pm

Reception in “le Grand Salon”

Tuesday, May 14, 2024

9:00 pm

Disco in “le Grand Salon”

Wednesday, May 15, 2024

7:30 pm

Barbecue in the yard

Live music “Brass is Here”

Stargazing in the Austran sky
with a former astronomer from Grenoble University - Yard and/or Salle Vercors

Programme overview

13th International Seminar on Speech Poduction. 13 -17 May 2024

	Monday May 13	Tuesday May 14	Wednesday May 15	Thursday May 16	Friday May 17
08:00am					
08:30am		**Caroline Niziolek**	**Sophie Scott**	**Maria Florencia Assaneo**	**Jason A. Shaw**
09:00am					
09:30am		Oral Session 1 **Feedback**	Oral Session 4 **Adaptation II**	Oral Session 7 **Coordination I**	Oral Session 11 **Coordination II**
10:00am					
10:30am		Coffee Break	Coffee Break	Coffee Break	Coffee Break
11:00am		Oral Session 2 **Adaptation I**	Oral Session 5 **Coarticulation**	Oral Session 8 **Phonetics/Phonology II**	Poster Session 4
11:30am					
12:00am		Oral Session 3 **Production/Perception**	Oral Session 6 **Phonetics/Phonology I**	Oral Session 9 **Development**	
12:30am					
01:00pm		Lunch Break	Lunch Break/free time	Lunch Break	Lunch break
01:30pm					
02:00pm					End of the Conference
02:30pm					
03:00pm		Poster Session 1		Poster Session 3	
03:30pm					
04:00pm			Coffee Break		
04:30pm			Poster Session 2		
05:00pm		Coffee Break		Coffee Break	
05:30pm		**Doris Mücke**		Oral Session 10 **Methodology**	
06:00pm					
06:30pm	Welcome Check-in		**Adrien Meguerditchian**		
07:00pm					
07:30pm	Aperitif/Dinner	Dinner	Dinner	Dinner	
08:00pm					

Keynote lectures

Tuesday, May 14, 2024

8:30 – 9:30 am

Caroline Niziolek
Communication Sciences and Disorders, University of Wisconsin–Madison, USA - (homepage)

Sensorimotor learning as a window to speech planning

How are speech movements planned? Typically, speech production is conceptualized as having separate linguistic and motor planning stages: psycholinguistic models select abstract units (e.g., phonemes or syllables), and models of speech motor control "read out" these units into articulatory movements. However, there is growing evidence that phonemic or syllabic motor programs alone are insufficient to explain patterns of speech behavior, necessitating models in which higher-level linguistic context is incorporated into the motor planning process. In this talk, I address the scope of speech planning through a series of experiments that use auditory feedback errors to induce learned changes to the pronunciation of speech sounds. This learning can occur in a context-specific manner, with speakers differentially changing their production of the same phoneme in opposite directions based on its word context. Here, we use sensorimotor learning as a marker of the influence of linguistic context, assessing whether adaptive changes can be differentiated by lexical context, syllable position, suprasegmental pitch, and word meaning. The results of these studies delineate when multisyllabic speech is planned holistically and when it relies on pre-specified motor programs that are sequenced online.

5:30 – 6:30 pm

Doris Mücke
IfL-Phonetikcs, University of Cologne, Germany - (homepage)

Multidimensionality of prosodic prominence: From neurotypical to atypical speech patterns

To overcome limitations imposed by symbolic approaches, researchers from many disciplines have turned to the framework of dynamical systems describing a multitude of different cognitive processes including the production and perception of speech sounds and their cognitive representations as well as movement coordination. One potential strength of dynamical systems is that they can handle a high amount of variability, because they do not separate between discrete symbolic representations and the continuous representations of the physical world. We will discuss the application of dynamical systems to capture prominence modulations of the speech system and its relation to linguistic functions on a multidimensional scale including intonational and textual variation. We will show how acoustic and articulatory modulations can change in relative importance with respect to prominence cuing in highly flexible way. Further, multidimensionality will be extended to multimodality of prosodic prominence, including co-speech head gestures from a dynamical perspective in different speaking styles. We conclude with the application of dynamical systems to impaired speech (Parkinson's disease). Speakers aim to compensate for problems of the speech motor system in a multidimensional phonetic space, which can be difficult to capture. In this respect, automatic acoustic speech analysis may be a promising tool to capture speech changes in speech disorders on a multidimensional scale.

Wednesday, May 15, 2024

8:30 – 9:30 am

Sophie Scott

Institute of Cognitive Neuroscience, University College London, UK - (homepage)

What's in a voice - from neural mechanisms to social influences

In this talk I will explore the implications of the fact that when we hear someone speaking, we also always hear a voice. I will map out the different kinds of information that are expressed in voices, and the ways that this interacts with spoken language. I will explore these interactions in both perception and production, and address some of the candidate neural systems that are recruited when speaking voices are heard and produced.

6:30 – 7:30 pm

Adrien Meguerditchian

CRPN, CNRS/Université Aix-Marseille, Marseille - (homepage)

The Gestural Origin of Language Production: Insight from the baboons' hands & brain specialization

Language is an unique communicative system involving hemispheric lateralization of the brain. To discuss the question of its origins, I will highlight the works on the communicative gestures in our primate cousins and their brain correlates. Indeed, nonhuman primates communicate mostly communicate not only with a rich vocal repertoire but also with manual and body gestures. In the last 20 years, we investigated this gestural system in the baboons Papio anubis, an Old World monkey species, as well as its lateralization and cortical correlates across development, using both ethological, psychology and longitudinal noninvasive in vivo brain imaging approach (MRI). In the present talk, I will summarize our main findings showing similar key intentional, referential "domain general" properties of language as well as some similar underlying structural hemispheric specialization including Broca, the Planum Temporale and the STS. I will also present our recent MRI longitudinal work documenting their brain ontogeny from birth and how they pave the way for the further emergence of gesture lateralization across development.

Thursday, May 16, 2024

8:30 – 9:30 am

Florencia Assaneo

Laboratorio de Percepción y Producción del Habla,
Instituto de Neurobiología, Universidad Nacional Autónoma de México, México - (homepage)

Causes and consequences of the syllabic rhythms

The speech signal is characterized by a rhythmic pattern of amplitude fluctuations, forming cycles composed of peaks and valleys. Surprisingly, these cycles, approximating the syllabic unit, exhibit temporal regularity across languages, typically oscillating between 3 and 6 cycles per second. This temporal regularity is not only present in the production of speech but also during its perception. It has been shown that when listening to speech, brain activity originating from auditory regions recovers the amplitude fluctuation of the perceived signal. In this presentation, I will discuss a series of studies delving into the interplay between the produced and perceived syllabic rhythm. Through our findings, I will present evidence supporting the hypothesis that the observed temporal regularity across languages may arise as a consequence of the underlying neural architecture supporting speech.

Friday, May 17, 2024

8:30 – 9:30 am

Jason A. Shaw

Department of Linguistics, Yale University, USA - (homepage)

Intentional dynamics in speech production

Speech production, like controlled actions more generally, involve selecting movement parameters from a continuous range of possibilities. In this talk, I consider how the dynamics of this cognitive process, which I refer to as intentional dynamics, relate to patterns of variability observed in speech. I formalize the dynamics using the tools of Dynamic Field Theory, treating the parameters of gesture control as the dimensions of Dynamic Neural Fields (DNFs). The fields evolve over time forming activation peaks under the influence of multiple excitatory and inhibitory forces. Formalized in this way, we can understand a number of well-known effects in speech production, including trace effects in speech errors, contrastive hyper-articulation, phonetic convergence/divergence to an interlocuter, and incomplete neutralization, as natural consequences of the intentional dynamics underlying cognitive control of speech.

List of abstracts

Day I

Tuesday, May 14

08:00am	
08:30am	**Caroline Niziolek**
09:00am	
09:30am	Oral Session 1
10:00am	**Feedback**
10:30am	Coffee Break
11:00am	Oral Session 2
11:30am	**Adaptation I**
12:00am	Oral Session 3
12:30am	**Production/Perception**
01:00pm	Lunch Break
01:30pm	
02:00pm	
02:30pm	

03:00pm	Poster Session 1
03:30pm	
04:00pm	
04:30pm	
05:00pm	Coffee Break
05:30pm	**Doris Mücke**
06:00pm	
06:30pm	
07:00pm	
07:30pm	Dinner
08:00pm	

Oral session 1

Feedback

9:30 – 10:30 am

	Title	Authors
9:30 - 9:50 am	Investigating the Effects of Auditory and Somatosensory Feedback on Laryngeal and Articulatory Motor Control in Individuals with Parkinson's Disease	Hasini R Weerathunge (Boston University)*; Nicole Tomassi (Boston University Stepp Lab); Daria Dragicevic (Boston University); Courtney Dunsmuir (Boston University); Megan Cushman (Boston University); Taylor Feaster (Boston University); Defne Abur (University of Groningen); Frank H. Guenther (Boston University); Cara Stepp (Boston University)
9:50 - 10:10 am	Somatosensory and visual influences on the perception of high rounded vowels	Jian-zhi Huang (National Taiwan University)*; Chenhao Chiu (National Taiwan University)
10:10 - 10:30 am	Compensatory response to tongue perturbation occurs similarly with normal and altered auditory feedback	Morgane Bourhis (GIPSA-Lab)*; Yosra Jelassi (GIPSA-Lab); Christophe Savariaux (GIPSA-lab); Pascal Perrier (Gipsa-lab, Grenoble INP, Université Grenoble Alpes); Takayuki Ito (GIPSA-lab)

Investigating the Effects of Auditory and Somatosensory Feedback on Laryngeal and Articulatory Motor Control in Individuals with Parkinson's Disease

Hasini R. Weerathunge[1,2], Courtney J. Dunsmuir[2,3], Daria A. Dragicevic[2], Nicole E. Tomassi[2,4], Megan R. Cushman[2], Taylor F. Feaster[2], Defne Abur[7,8], Frank H. Guenther[1,2], Cara E. Stepp[1,2, 6]*

[1]Department of Biomedical Engineering, Boston University, Boston, MA
[2]Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA
[3]Department of Occupational Therapy, Boston University, Boston, MA
[4]Graduate Program for Neuroscience, Boston University, MA
5Department of Otolaryngology-Head and Neck Surgery, Boston University School of Medicine, Boston, MA
[6]Department of Otolaryngology-Head and Neck Surgery, Boston University School of Medicine, Boston, MA
[7]Department of Computational Linguistics, University of Groningen, the Netherlands
[8]Research School of Behavioral and Cognitive Neurosciences, University of Groningen, the Netherlands

hasiniw@bu.edu, dunsmuir@bu.edu, ddragic@bu.edu, ntomassi@bu.edu, tfeast@bu.edu, d.abur@rug.nl, guenther@bu.edu, cstepp@bu.edu

Introduction. Idiopathic Parkinson's disease (PD) is the fastest growing neurodegenerative disease in the world, and has detrimental effects on motor and non-motor control, as well as sensory function (Dorsey et al., 2018). Approximately 90% of persons with PD (PwPD) develop hypokinetic dysarthria, a motor speech disorder that presents as various deficits in speech production as manifested in vocalization and articulation (Ho et al., 1998). Identifying the pathophysiology of speech motor control deficits in PwPD can lead to better clinical intervention and management of speech dysfunction in this population. Prior research has identified impaired sensorimotor learning capabilities and higher reliance on auditory feedback for speech motor control in PwPD compared to typical age- and sex- matched adults (Abur et al., 2018; Kiran & Larson, 2001; Mollaei et al., 2016; Mollaei et al., 2013). However, the role of somatosensory feedback in PD for speech production has not been explored. This study aims to investigate the differential contributions of auditory and somatosensory feedback control mechanisms of laryngeal and articulatory speech subsystems in PwPD. In this study, we investigated how laryngeal and articulatory motor control is differentially affected by PD by examining sensorimotor measures at the group level for PwPD compared to control speakers. Altered feedback paradigms are commonly used to examine underlying auditory-motor function of groups with or without motor speech disorders. Reflexive paradigms include sudden and unpredictable perturbations of sensory feedback, and are used to investigate sensory feedback-based error correction capabilities of laryngeal and articulatory speech subsystems. Adaptation paradigms include gradual and predictable perturbations of sensory feedback, and are used to investigate auditory-motor integration capabilities (i.e., the ability to update preexisting motor programs based on persistent feedback-based errors).

Methods. Thirty-four PwPD (16 females, 18 males; age = 67 ∓ 8 years, 53 - 80 years) and 34 age-and sex-matched speakers with typical speech (age = 67 ∓ 8 years, 50 – 81 years) were enrolled in the study. PwPD were on their typical PD medication schedule while being tested. A series of sensorimotor measures were conducted related to auditory and somatosensory feedback error correction (i.e., via auditory and somatosensory reflexive altered feedback paradigms) and auditory-motor integration (i.e., via auditory adaptive altered feedback paradigms). Participants produced the consonant-vowel-consonant words *bid, hid,* and *id* in 60-trial paradigms, each providing a specific sensorimotor measure. Predictable auditory perturbations were applied to vocal fundamental frequency (i.e. vocal f_o) or vowel first formant (i.e., vowel F_1) to extract vocal f_o and vowel F_1 ***auditory adaptive responses***, respectively. Sudden and unpredictable auditory perturbations were applied to vocal f_o or vowel F_1 to extract vocal f_o and vowel F_1 ***auditory reflexive responses***, respectively. Physical perturbations were applied via a small, tubular, inelastic balloon, constructed with heavy-duty nitrile material to the larynx (i.e., applying superior-posterior pressure on the anterior neck at the level of the thyroid cartilage) or the jaw (i.e., applying inferior pressure on the jaw via the lower molars). The acoustic consequences produced during physical perturbations (i.e., in vocal f_o for laryngeal perturbations and in vowel F_1 for jaw perturbations) was masked with speech-shaped noise, with the objective of measuring the isolated ***somatosensory reflexive responses*** of participants to somatosensory feedback variations generated by the physical perturbations. We did not expect to observe statistically significant variations in auditory adaptive responses of vocal f_o and vowel F_1 for PwPD compared to controls based on prior research conducted on PwPD on typical medication (Abur et al., 2021). We expected to observe higher auditory reflexive response for vocal f_o and lower auditory reflexive response for vowel F_1 in PwPD compared to controls based on prior research (Abur et al., 2021; Mollaei et al., 2016). We expected reduced somatosensory reflexive responses in the group with PD based on prior research suggesting reduced somatosensory function in PD (Conte et al., 2013; Hammer & Barlow, 2010). Mixed methods analyses of variance (ANOVAs) were calculated with group (i.e., PD, Control) and adaptation phase (i.e., Baseline, Hold1, Hold2, Aftereffect) as fixed factors for vocal f_o and vowel F_1 auditory adaptive responses. Four one-tailed two-sample t-tests were calculated to identify statistically significant differences between the groups for vocal f_o and vowel F_1 auditory reflexive response, and laryngeal and jaw somatosensory reflexive responses.

Results. Data from 30 PwPD (15 females, 15 males; age M = 65.9, SD = 7.4 years) and 30 age – and sex- matched speakers (age M = 66.4, SD = 7.8 years) with typical speech function were included in the preliminary analysis. We anticipate completion of the data analysis of full dataset (i.e., 34 participants in each group) at the time of the presentation. Preliminary results show statistically significantly lower opposing responses to somatosensory feedback in the laryngeal domain in the group with PD (t = -2.12; p = .020). The effect size for the difference between the groups was calculated using Cohen's d, resulting in a value of 0.57, which is considered a medium effect. The findings suggest that somatosensory feedback control mechanisms may be impaired in PwPD, specifically in the laryngeal speech production subsystem. There were no significant differences in responses observed between PwPD and controls for auditory perturbation responses for vocal f_o or vowel F_1. However, the group means indicate that the results were in opposing direction of the directional hypothesis for auditory reflexive responses for vocal f_o. Similarly, there were no significant differences in responses for somatosensory perturbations of the jaw. **Tables 1** and **2** show statistical analysis results for the mixed methods ANOVAs for adaptive responses and t-tests for reflexive responses, respectively.

Table 1. *Mixed methods analyses of variance on adaptive response magnitudes.*

Response Type	Effect	*df*	η_p^2	*F*	*p*
Auditory Adaptive vocal f_o	Group	1	0.00	0.69	.407
	Adaptation Phase	3	0.01	1.05	.370
	Group * Phase	3	0.00	0.12	.950
Auditory Adaptive vowel F_1	Group	1	0.00	0.31	.580
	Adaptation Phase	3	0.00	0.38	.764
	Group * Phase	3	0.01	0.14	.936

*Significant at p < .05; significant p-values bolded and marked with *; η_p^2 effect sizes: small (<.06), medium (.06–.14), large (>.14);

Table 2. *One-tailed two-sample t-tests on reflexive response magnitudes.*

Response Type *(unit)*	PD		Control		*t*	*p*	*Cohen's d*
	M	*SD*	*M*	*SD*			
Auditory Reflexive vocal f_o *(cents)*	-8.2	18.3	-16.4	16.6	t(43) =1.58	.06	0.47
Auditory Reflexive vowel *F*1 *(percent)*	0.45	4.40	0.29	5.68	t(46) = 0.11	.545	0.03
Somatosensory reflexive laryngeal	0.77	1.16	1.35	0.84	t(45) = -2.12	**.020***	0.57
Somatosensory reflexive jaw	1.06	1.87	0.76	0.67	t(33) = 0.78	.780	0.21

*Significant at p < .05; Significant p values bolded and marked with *. Cohen's d effect sizes: 0.2 = small, 0.5 = medium, 0.8 = large;

Discussion. This is the first study to comprehensively investigate the contributions of auditory and somatosensory feedback control mechanisms and the contributions of articulatory and laryngeal subsystems of speech in PwPD. The study results indicate that there are detrimental effects of PD on somatosensory control of the larynx. These results also provide evidence that the laryngeal and articulatory speech production subsystems operate with differential auditory and somatosensory feedback control mechanisms. In combination with previous work, the outcomes further suggest that current models of speech motor control should consider decoupling laryngeal and articulatory domains to better model speech motor control processes. The study outcomes will be instrumental in enhancing clinical interventions on PwPD to target affected speech subsystems and feedback control mechanisms.

References

Abur, D., Lester-Smith, R. A., Daliri, A., Lupiani, A. A., Guenther, F. H., & Stepp, C. E. (2018). Sensorimotor adaptation of voice fundamental frequency in Parkinson's disease. *PloS one, 13*(1), e0191839. https://doi.org/10.1371/journal.pone.0191839

Abur, D., Subaciute, A., Daliri, A., Lester-Smith, R. A., Lupiani, A. A., Cilento, D., Enos, N. M., Weerathunge, H. R., Tardif, M. C., & Stepp, C. E. (2021, Dec 13). Feedback and Feedforward Auditory-Motor Processes for Voice and Articulation in Parkinson's Disease. *J Speech Lang Hear Res, 64*(12), 4682-4694. https://doi.org/10.1044/2021_JSLHR-21-00153

Conte, A., Khan, N., Defazio, G., Rothwell, J. C., & Berardelli, A. (2013, Dec). Pathophysiology of somatosensory abnormalities in Parkinson disease. *Nat Rev Neurol, 9*(12), 687-697. https://doi.org/10.1038/nrneurol.2013.224

Dorsey, E., Sherer, T., Okun, M. S., & Bloem, B. R. (2018). The emerging evidence of the Parkinson pandemic. *Journal of Parkinson's disease, 8*(s1), S3-S8.

Hammer, M. J., & Barlow, S. M. (2010, Mar). Laryngeal somatosensory deficits in Parkinson's disease: implications for speech respiratory and phonatory control. *Exp Brain Res, 201*(3), 401-409. https://doi.org/10.1007/s00221-009-2048-2

Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1998). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural neurology, 11*(3), 131-137.

Kiran, S., & Larson, C. R. (2001, Oct). Effect of duration of pitch-shifted feedback on vocal responses in patients with Parkinson's disease. *J Speech Lang Hear Res, 44*(5), 975-987. https://doi.org/10.1044/1092-4388(2001/076)

Mollaei, F., Shiller, D. M., Baum, S. R., & Gracco, V. L. (2016, Sep 1). Sensorimotor control of vocal pitch and formant frequencies in Parkinson's disease. *Brain Res, 1646*, 269-277. https://doi.org/10.1016/j.brainres.2016.06.013

Mollaei, F., Shiller, D. M., & Gracco, V. L. (2013, Oct). Sensorimotor adaptation of speech in Parkinson's disease. *Mov Disord, 28*(12), 1668-1674. https://doi.org/10.1002/mds.25588

Somatosensory and visual influences on the perception of high rounded vowels

Jian-zhi Huang[1], Chenhao Chiu[1,2,3]

[1] Graduate Institute of Linguistics, National Taiwan University
[2] Graduate Institute of Brain and Mind Sciences, National Taiwan University
[3] Neurobiology and Cognitive Science Center, National Taiwan University
r10142007@ntu.edu.tw, chenhaochiu@ntu.edu.tw

Introduction. Speech perception is multimodal (Keough et al., 2019; Rosenblum, 2008). Studies have found that speech perception is influenced not only by vision (McGurk effect: McGurk & MacDonald, 1976; lip-reading of roundedness: Trudeau-Fisette et al., 2022), but also by our somatosenses (aerotactile feedback: Gick & Derrick, 2009; proprioception to perturbation: Ito et al., 2009; corollary discharge from inner speech: Scott et al., 2013). This evidence shows that speakers do not entirely rely on acoustic feedback in speech communication, but these multimodal feedback mechanisms are integrated with each other. For example, results from Scott et al. (2013) show that inner speech, including mouthing and imagining, guides our auditory perception to the target segments being mouthed or imagined; it also shows a comparable effect when we perform inner speech on the sub-phonemic contents that share the same place of articulation. Similarly, results from Trudeau-Fisette et al. (2022) show that visual feedback also guides our auditory perception. When participants are visually prompted with rounded vowels, they tend to identify the perceived auditory stimuli as a rounded vowel, and vice versa for unrounded vowels. Crucially, even in the absence of acoustics, these multimodal feedback mechanisms still interact with each other. Masapollo & Guenther (2019) demonstrated that the insertion of a lip tube to create more extreme [u] gestures enhanced native English speakers' ability to discriminate between cross-language lip postures in [u] videos (English and French, in their case). This suggests that somatosensory feedback not only influences but also modulates our visual perception. Information encoded in these multimodal feedback systems appears to be phonemic (Scott et al. 2013), and phonemic contrasts are rooted in distinctive features present in the language. In the same study, Scott et al. (2013) also proposes that feedback not only encompasses phonemic information but also incorporates sensory information, echoing the findings from Masapollo & Guenther (2019) that somatosensory feedback provides subtle sensory information about the visual lip postural difference between English [u] and French [u]. Thus, subtle sensory difference can also be affected by multimodal feedback. Building on the implication from Masapollo and Guenther (2019), it shows that not all features classified as [+round] exhibit identical visual characteristics. The sensitivity to distinctions between the two rounded vowels in videos can be enhanced through somatosensory feedback. Therefore, our study aims to delve into the specific case of the two high rounded vowels in Taiwan Mandarin—high back rounded vowel /u/ and high front rounded vowel /y/. Despite that both [u] and [y] are associated with [+round], [u] exhibits a more circular round posture whereas [y] exhibits a more laterally compressed posture (Chiu & Huang, 2023). It is of research interest to determine whether multimodal information can alter the perception of sounds that contrast in lip and tongue postures. In this vein, the present study investigates whether the lip postural difference in Taiwan Mandarin high rounded vowels can be modulated by multimodal feedback.

Methods. The stimuli comprised two sets of 11-step [u] to [y] continua generated through *STRAIGHT* (Kawahara et al., 2008), utilizing natural productions of [u] and [y] by two gender-balanced talkers with distinctive /u/ vs. /y/ lip postures. Participants were all native Taiwan Mandarin speakers with no hearing or visual impairments. They were asked to perform a forced-choice task identifying acoustic [u] and [y] in three between-subject experiment conditions: inner speech, visual-only, inner speech + visual. The procedures for the three experimental conditions were identical: pre-test, feedback conditions, and post-test. Both the pre- and post-tests were conducted under auditory-only conditions. In the feedback condition of Experiment 1, participants were asked to perform two types of inner speech (mouthing and imagining) while identifying the auditory stimulus from an 11-step /u-y/ continuum. In the feedback condition of Experiment 2, the auditory stimuli from the /u-y/ continuum were synchronized with visually articulated [a, u, y]. In the feedback condition of Experiment 3, participants were instructed to engage in inner speech on target vowels [u, y] while the auditory stimuli were synchronized with visually articulated [u, y] both congruently and incongruently. Inner speech in Experiment 1 and Experiment 3 was synchronized by a three-second countdown prior to the presentation of speech stimuli. In the first two experiments, vowel [a] was included as a control to discern the effects of the lips from the tongue. For data analyses, mixed-effects logistic regression modeling was run on the responses of [u] and [y], coded as 0 and 1, respectively.

Results. Figure 1(a) shows the effect of mouthing from Experiment 1. Compared to the pre-test baseline, mouthing [u] and mouthing [y] both influence participants' perception—mouthing [u] triggers a stronger perceptual shift towards [u] ($\beta = -0.764$, $p < .05$), and mouthing [y] triggers a stronger perceptual shift towards [y] ($\beta = 1.997$, $p < .001$). No effects are observed for mouthing [a] ($\beta = -0.322$, *N.S.*) or the post-test ($\beta = 0.215$, *N.S.)*. A similar but smaller effect of imagining was found, compared with mouthing. When provided with visually articulated [u] and [y] (Experiment 2, as in Figure

1(b)), participants shifted their responses towards [u] (β = -1.084, $p < .001$) and [y] (β = 1.795, $p < .001$) when visual and auditory information congruently matched with each other. No effect of visually articulated [a] (β = -0.105, *N.S.*) was found, and neither was the post-test (β = 0.449, *N.S.*).

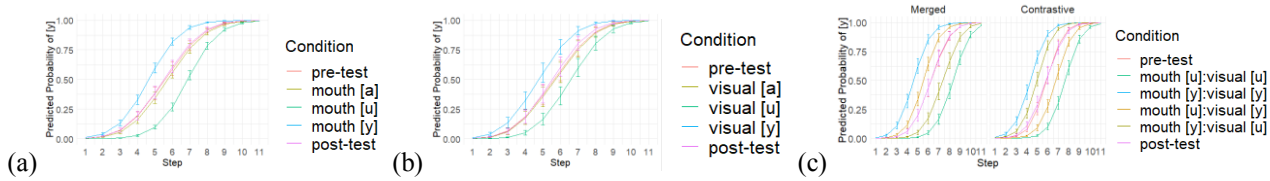


Figure 1: *Probability of [y] identification in (a) Experiment 1: mouthing, (b) Experiment 2: vowel videos, and (c) Experiment 3: mouthing: vowel videos. Results were separated by those who made lip postural contrasts between [u] and [y] during mouthing ('Merged') and those who didn't ('Contrastive').*

The results from Experiment 3 (mouthing + visual) are shown in Figure 1(c). As revealed, when the provided feedback in both modalities, i.e., mouthing and visual, are matched, significant perceptual shifts toward the matched vowels are observed, e.g., mouthing [u] with visual [u] towards [u] (β = -4.481, $p < .001$) and mouthing [y] with visual [y] towards [y] (β = 5.428, $p < .001$). However, when the provided feedback in both modalities did not match, two types of responses are observed according to the participants' own productions of the two rounded vowels, i.e., an interaction of the participants' own lip postural contrast and feedback conditions. When participants make no lip postural difference between [u] and [y] during mouthing (the *Merged*, baseline in mixed-effects logistic regression modeling), the perceptual shift prioritizes the visual modality, i.e., mouthing [u] with visual [y] yielded more [y] responses (β = 1.967, $p < .01$) and mouthing [y] with visual [u] yielded more [u] responses (β = -2.483, $p < .001$). On the other hand, for those who make lip postural contrasts between [u] and [y] during mouthing (the *Contrastive*), the shift prioritizes the somatosensory modality, i.e., mouthing [u] with visual [y] attracted more [u] responses (β = -3.757, $p < .001$) and vice versa for [y] (β = 4.451, $p < .001$). Post-tests report no effect (β = 0.242, *N.S.*). Similar but smaller effects of imagining were also found compared with mouthing.

Discussion.

The study investigates the influence of somatosensory and visual prompts on high rounded vowel perception. Our results showed that both mouthing and imagining the target vowel [u, y] influence the auditory perception, suggesting that the perception of these two rounded vowels are modulated by somatosensory feedback (Exp. 1) and visual prompts (Exp. 2) in addition to tongue position. The results also suggest that somatosensory feedback may lie beyond the level of phoneme. The integration between somatosensory and visual information was also explored (Exp. 3). Matched somatosensory-visual feedback prompts a shift in responses while mismatched somatosensory-visual feedback results in a modality preference based on the participants' own production.

In conclusion, lip postural difference of the two rounded vowels can be captured through somatosensory and visual feedback. Furthermore, the efficacy of somatosensory and visual feedback is enhanced when the integrated feedback is congruent, but contingent upon individuals' production behaviors when feedback is incongruent. These findings increase our understanding of the multimodal nature of speech perception and its potential of being influenced by production experience.

References

Chiu, C., & Huang, P-S. (2023). Lip postures of high vowels in Taiwan Mandarin. In: Radek Skarnitzl & Jan Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1052–1056). Guarant International.

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, *462*(7272), 502–504.

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences*, *106*(4), 1245–1248.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3933–3936). IEEE.

Keough, M., Derrick, D., & Gick, B. (2019). Cross-modal effects in speech perception. *Annual review of linguistics*, *5*, 49–66.

Masapollo, M., & Guenther, F. H. (2019). Engaging the articulators enhances perception of concordant visible speech movements. *Journal of Speech, Language, and Hearing Research*, *62*(10), 3679–3688.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current directions in psychological science*, *17*(6), 405–409.

Scott, M., Yeung, H. H., Gick, B., & Werker, J. F. (2013). Inner speech captures the perception of external speech. *The Journal of the Acoustical Society of America*, *133*(4), EL286-EL292.

Trudeau-Fisette, P., Arnaud, L., & Ménard, L. (2022). Visual Influence on Auditory Perception of Vowels by French-Speaking Children and Adults. *Frontiers in Psychology*, *13*, 740271.

Compensatory response to tongue perturbation occurs similarly with normal and altered auditory feedback

Bourhis M., Jelassi Y., Savariaux C., Perrier P., Ito T.
Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab
Morgane.bourhis@grenoble-inp.fr, takayuki.ito@grenoble-inp.fr

Introduction: In a former study (Ito et al 2020), using a sudden force perturbation, we have provided clear evidence for a quick compensatory response (around 130ms latency) in the tongue for posture stabilization during vowel production. This study was complemented by Bourhis et al, (2022) who have demonstrated that auditory masking does not alter this compensatory response, suggesting that somatosensory feedback could be the main sensory source of this response. In the current study, we want to further examine the possible contribution of auditory feedback in tongue stabilization mechanisms. We combined the tongue perturbation with a simultaneous alteration of the auditory feedback based on real-time formant shift. Our prediction was that the latency of the auditory-based compensation should be longer than the somatosensory one, and hence that auditory feedback alteration should not affect the timing if the quick compensatory response induced by the tongue perturbation. The compensation for the auditory error should occur in a period later than that of the compensatory response due to the somatosensory error.

Method: 12 native French speakers with no history of auditory impairment participated in the experiment. They were asked to sustain vowel /ɛ/ for 3 s in response to the appearance of a visual cue. Vowel production started and ended with closed mouth. We simultaneously recorded articulatory displacements of tongue and jaw using electromagnetic articulography (Wave, Northern Digital Inc.) and speech acoustics using Audapter (Cai et al, 2011). The speech signal was sampled at 22 kHz and articulatory movements at 200 Hz. Six sensors were placed on the upper lip, lower lip, jaw, tongue tip, blade and dorsum in the mid-sagittal plane of the head. Reference sensors were also placed on the nasion, left and right mastoids, and the upper incisor for head movement correction. In the test, we applied two perturbations, namely the tongue mechanical perturbation and the auditory feedback perturbation. The tongue perturbation was applied with a small robotic device (Phantom Premium 1.0, Geomagic) that was connected to the tongue through a thin thread glued on both lateral sides of the tongue blade. During mechanical perturbations a 1N force pulled the tongue forward. The auditory perturbation was applied using Audapter: formant F1 was shifted by 20% either upwards or downwards, and the altered sound was played back through magnetic compatible earphones (Natus Tip 300). These two perturbations were applied for 1 s after the onset of the vocalization. Because of a technical difficulty to perfectly synchronize the onsets of these two perturbations, auditory perturbation occurred slightly in advance of the tongue perturbation when both were applied together.
We tested five perturbed conditions combining altered auditory feedback and tongue perturbation: altered auditory feedback alone (AAFup and AAFdown), tongue perturbation alone (PTB), and altered auditory feedback with tongue perturbation (AAFup+PTB and AAFdown+PTB). In total, 225 trials were carried out. The perturbation was applied in the pseudo randomly selected one third of trials. All five perturbed conditions were applied every 15 trials. In total, 15 responses were recorded per condition.

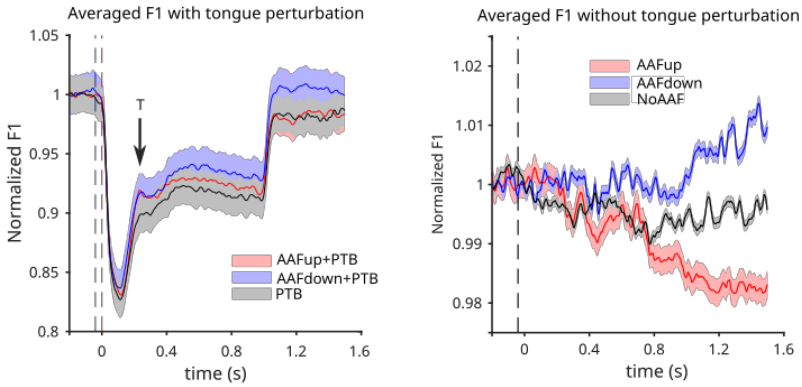


Figure 1: Left panel: F1 variation associated with auditory perturbations alone; Right panel: F1 responses to the tongue perturbation under the three auditory conditions.

We focus here on the analysis of the acoustical data. F1 values were estimated using linear predictive coding using 20 ms time windows shifted at a 10 ms rate. Trials with wrong estimation (F1<300 Hz or F1>700 Hz) were removed from the analysis. Two participants were also removed from the analysis due to high trial-to-trial variability. For each trial, time zero was set at the onset of the tongue perturbation, and the thus aligned within-trial F1 time variations were averaged across perturbed trials in each condition and in each participant. Because of its large variability across trials, F1 was normalized via the division by is baseline amplitude, which is defined as the mean of the 50 ms interval preceding the auditory perturbation. We measured the time location T of the local maximum that corresponds to the peak of the compensatory response (Figure 1 right panel, arrow 'T'), due to the combination of passive and somatosensory effects (Ito et al 2020). A repeated measure one-way ANOVA was applied across the conditions for the statistical analysis of T. In order to examine the latency of the compensatory response based on auditory error we compared two auditory perturbation conditions in two pairs (1: AAFup+PTB and AAFdown+PTB, and 2: AAFup and AAFdown).

Results: We first compared the F1 responses to the perturbation under the different auditory conditions when tongue perturbation was applied (PTB, AAFup+PTB and AAFdown+PTB). They were all similar with in particular a change in the slope of the F1 variation at around 200 ms (no significant difference across auditory condition at time T, $p > 0.05$). In the two conditions combining the mechanical and the auditory perturbation (AAFup+PTB and AAFdown+PTB), we also observed visually a small deviation between the two F1 variations at around 400ms, although there was no significant difference. This small deviation may correspond to the compensation for the auditory perturbation. To verify this idea, we also compared the two conditions with altered auditory feedback alone (AAFup and AAFdown). We then observed a discrepancy between the two F1 variations from around 400 ms after the onset of the perturbation (see Figure 1, left panel). However, differences between normal and altered auditory feedback conditions seem to differ after 500ms. This indicates that a compensation in response to the auditory perturbation was also induced, but with a latency significantly longer than the one induced by the tongue perturbation.

Discussion: No reliable differences were found between auditory conditions in the F1 response to the mechanical tongue perturbation. This clearly supports our hypothesis that somatosensory feedback is the main source of the quick compensatory response for posture stabilisation. We also found a divergence in the formant variation between the two altered auditory feedback conditions in both pairs (AAFup+PTB vs AAFdown+PTB and AAFup vs AAFdown). This divergence occurred with a latency that is clearly longer (>400 ms) than the one of the somatosensory-based compensatory response (200 ms). Cai et al (2011) showed relatively shorter latency in the compensatory response to auditory perturbation (around 130 ms), but it was in a dynamic speech production task, associated with triphtongs production, The larger latency observed in our study could be due to the static aspect of our speech production task. This is supported by Purcell & Munhall (2006) who showed a relatively long latency (> 400 ms) in the compensatory response to single exposure to auditory perturbation in a sustained vowel production task similar to the one used in the current study. Although it is still unknown why this difference occurs, it seems that static tasks generate longer latencies in auditory compensatory mechanisms than dynamic ones. Overall, our results provide additional evidence that somatosensory feedback enables faster compensatory responses than auditory feedback does. This supports our hypothesis that somatosensory-based reflex mechanism plays a major role in tongue posture stabilisation during the vowel production.

References:

Bourhis, M., Perrier, P., Savariaux, C., Ito, T. (2022). Compensatory movement of the tongue for speech production with or without masking noise. SMC 2022 - 8th International Conference on Speech Motor Control, Aug 2022, Groningen, Netherlands. ⟨hal-03878063⟩

Cai, S., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2011). Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. Journal of Neuroscience, 31(45), 16483–16490. https://doi.org/10.1523/JNEUROSCI.3653-11.2011

Ito, T., Szabados, A., Caillet, J. L., & Perrier, P. (2020). Quick compensatory mechanisms for tongue posture stabilization during speech production. Journal of Neurophysiology, 123(6), 2491–2503.

Purcell, D. W., & Munhall, K. G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America*, *120*(2), 966–977. https://doi.org/10.1121/1.2217714

Tremblay S, Shiller DM, Ostry DJ (2003) Somatosensory basis of speech production. Nature 423:866–869.

Oral session 2

Adaptation I

11:00 – 12:00 am

	Title	Authors
11:00 - 11:20 am	Planning competing values of a single phonological feature vs. planning values for multiple features	Kevin D Roon (CUNY Graduate Center)*; D. H. Whalen (CUNY Graduate Center)
11:20 - 11: 40 am	Cerebellar degeneration eliminates adaptation to perturbations of segmental duration in speech	Robin Karlin (University of Missouri)*; Ben Parrell (University of Wisconsin-Madison)
11:40 - 12:00 am	Lingual motor learning based on real-time visual feedback in individuals with and without Parkinson's disease	Teja Rebernik (University of Groningen)*; Mark Tiede (Yale University); Thomas B Tienkamp (University of Groningen); Defne Abur (University of Groningen); Jidde Jacobi (University of Groningen); Martijn Wieling (University of Groningen)

Planning competing values of a single phonological feature vs. planning values for multiple features

Kevin D. Roon[1], D.H. Whalen[1,2,3]

[1]CUNY Graduate Center, Program in Speech-Language-Hearing Sciences
[2]Yale Child Study Center
[3]Yale University, Department of Linguistics
kroon@gc.cuny.edu, dwhalen@gc.cuny.edu

Introduction. In the domains of speech science and phonetics, the phrase "models of speech production" most often refers to models of how articulator movements are planned and/or controlled (Saltzman & Munhall, 1989; Tourville & Guenther, 2011), and the phenomena of interest are usually articulatory movements and their acoustic consequences. In psycholinguistics, "models of speech production" refers to models of how lexical items are selected from memory (e.g., Dell, 1986; Roelofs, 2000; Seidenberg & McClelland, 1989) for production, and the phenomena of interest are very often verbal response times (VRTs). Phonetic models rarely (if ever) consider VRTs, and psycholinguistic models rarely consider phonetic data. Furthermore, the only representations that are reliably shared between these two types of models are phonemes. One can get the impression from the literature that the models in these two domains have nothing to do with each other, while what is needed is greater refinement of models of both types to enable more integration. In this study, we focus on one important aspect of that goal. Specifically, we note that psycholinguistic models do not include representations more fine-grained than the phoneme, despite ample empirical evidence of such representations VRTs in a variety of experimental tasks (e.g., Gordon & Meyer, 1984; Mousikou et al., 2015).

Results from a response-distractor task have also yielded feature-level effects on VRTs. In this task, participants are told to produce simple consonant-vowel syllables (e.g., /pa/, /ta/, /da/) based on some visual cue. Very soon after the presentation of that cue, a distractor stimulus is presented. The (dis)similarity of the distractor to the target response has repeatedly been found to modulate the VRTs of the participants (e.g., Galantucci et al., 2009). Roon and Gafos (2015) found that VRTs were modulated by feature-level (dis)similarity between a response and an audio distractor: for congruent trials, when the distractor matched the response being planned on all features except voicing (e.g., /ta/-/*da*/) or primary oral articulator (e.g., /ta/-/*pa*/), VRTs were longer than when there was no distractor or a tone distractor, but shorter than on incongruent trials, when the response and distractor mismatched on both voicing and articulator (e.g., /ta/-/*ba*/).

There was also an unexpected and surprising aspect of the data from Roon and Gafos (2015). There were two experiments in that study, which had largely the same design, but one key difference. In each experiment, a given block consisted of two possible responses. Within the blocks of Experiment 1, the primary oral articulator of the response was always predictable but voicing was not (e.g., /ta/~/da/), while within the blocks of Experiment 2, voicing was predictable but primary oral articulator was not (e.g., /ta/~/pa/). Figure 1 shows the VRTs from that study within distractor condition, separated by the experiment in which they were produced. VRTs in the "Unknown articulator" experiment were notably shorter (by 43 ms on average) than in the "Unknown voicing" experiment. This difference in VRTs held across distractor conditions, as well as on trials where there was no distractor. There is no model of speech production from any domain that would—or could—predict this difference.

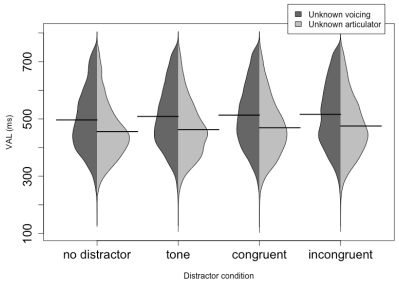


Figure 1: *Verbal response times from the two experiments in Roon and Gafos (2015).*

Roon and Gafos (2016) present a dynamical, computational model of phonological planning that accounts not only for the response-distractor compatibility results, but also for this cross-experiment difference in VRTs. The crucial notion in that model is that having to plan for two features that are inherently mutually exclusive incurs a processing cost that is not incurred when planning for two features that are not mutually exclusive. Having to plan for two values for voicing involves inherently mutually exclusive options: no single consonant can be voiced and voiceless. In contrast, constrictions

of primary oral articulators are not inherently mutually exclusive, witnessed by the fact that many sounds in many languages involve multiple concurrent constrictions made by different articulators.

A difference in VRTs > 40 ms suggests that it reflects an important and robust component of speech production. However, the empirical finding was unexpected, and its theoretical explanation was post-hoc. It was also the result of comparing VRTs from two different experiments that had different participants. Lastly, those differences were never subject to statistical analysis. The present study has 3 objectives: 1) to assess the cross-experiment VRT differences statistically, and to conduct a new experiment that will 2) replicate that result with a within-participant design, and 3) to further test the hypothesis that these differences are due to mutual exclusivity.

Methods. We assessed the statistical reliability of the data from Roon and Gafos (2015). The data from those two experiments were combined, with a new field (Experiment) added to each trial, either "Unknown voicing" or "Unknown articulator". To simplify the statistical analysis and to remove an possible influence of distractor, only trials on which there was no distractor or a non-linguistic tone distractor were included (leftmost two groups in Figure 1). 14605 trials were included in the analysis (7320 for Experiment "Unknown voicing", 7285 for Experiment "Unknown articulator").

A linear mixed-effect model of the log-transformed VRT data was created, which included random effects for participant and item (intercepts only); "control" fixed effects for SOA, previous trial log VRT, whether the response was the same as the previous trial; and the fixed effect of theoretical interest: Experiment.

The second part of our study is a new experiment that will attempt to replicate the finding from Roon and Gafos (2015), first with plosive-initial stimuli (as in the original) and then with fricative-initial stimuli. The experiment will also test the notion of mutual exclusivity by comparing VRTs when participants have to plan for conflicting values of tongue-tip constriction location (which are mutually exclusive) vs. when participants have to plan for constrictions of different primary oral articulators. Within a block, 50 English speaking participants will learn cue-response pairs (e.g., "If you see # #, say 'tuh', if you see & &, say 'duh'"). Our predictions will be tested by comparing responses sharing the same onset across combinations. The replication of Roon and Gafos (2015) will compare VRTs for /tʌ/ trials when the alternative response is /dʌ/ with those when the alternative response was /pʌ/. Next, VRTs for /sʌ/ trials when the alternative response is /zʌ/ will be compared with those when the alternative response was /fʌ/. Lastly, VRTs for /sʌ/ trials when the alternative response is /θʌ/ will be compared with those when the alternative response was /fʌ/. One lme model will be used for the comparison of /tʌ/ trials, and another for the comparison of /sʌ/ trials.

Results. The results of our statistical model showed a significant effect of Experiment (p = 0.015) showing that the differences shown in the left two groups of Figure 1 were reliable. Our predictions for VRTs for the planned experiment are: 1) $/tʌ/_{\text{alternative /dʌ/}} > /tʌ/_{\text{alternative /pʌ/}}$, 2) $/sʌ/_{\text{alternative /zʌ/}} > /sʌ/_{\text{alternative /fʌ/}}$, 3) $/sʌ/_{\text{alternative /θʌ/}} > /sʌ/_{\text{alternative /fʌ/}}$.

Discussion. The exploratory analysis above shows that the numeric differences in mean VRTs across the experiments from Roon and Gafos (2015) were statistically reliable, for responses whose onsets included coronal and velar, voiced and voiceless plosives as well as nasals. The results from the new experiment will clarify how the precise nature and details of those representations play a role in the models mentioned above. The implications of other combinations of outcomes will range from informing further refinement the model, to assessing the validity of the model and/or the appropriateness of the representations involved.

References

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Galantucci, B., Fowler, C. A., & Goldstein, L. M. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics*, *71*(5), 1138–1149. https://doi.org/10.3758/APP.71.5.1138

Gordon, P. C., & Meyer, D. E. (1984). Perceptual-motor processing of phonetic features in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(2), 153–178. https://doi.org/10.1037//0096-1523.10.2.153

Mousikou, P., Roon, K. D., & Rastle, K. (2015). Masked primes activate feature representations in reading aloud. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 636–649. https://doi.org/10.1037/xlm0000072

Roelofs, A. (2000). WEAVER++ and other computational models of lemma retrieval and word-form encoding. In L. R. Wheeldon (Ed.), *Aspects of Language Production* (pp. 71–114). Psychology Press. https://doi.org/10.4324/9781315804453

Roon, K. D., & Gafos, A. I. (2015). Perceptuo-motor effects of response-distractor compatibility in speech: beyond phonemic identity. *Psychonomic Bulletin & Review*, *22*(1), 242–250. https://doi.org/10.3758/s13423-014-0666-6

Roon, K. D., & Gafos, A. I. (2016). Perceiving while producing: Modeling the dynamics of phonological planning. *Journal of Memory and Language*, *89*, 222–243. https://doi.org/10.1016/j.jml.2016.01.005

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*(4), 333–382. https://doi.org/10.1207/s15326969eco0104_2

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. https://doi.org/10.1037/0033-295x.96.4.523

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952–981. https://doi.org/10.1080/01690960903498424

Cerebellar degeneration eliminates adaptation to perturbations of segmental duration in speech

Robin Karlin[1], Benjamin Parrell[2]

[1]Department of Speech, Language, and Hearing Sciences, University of Missouri
[2]Department of Communication Sciences and Disorders, University of Wisconsin – Madison
rkarlin@health.missouri.edu, bparrell@wisc.edu

Introduction. Cerebellar ataxia is a movement disorder caused by damage to or degeneration of the cerebellum. Temporal deficits, including both overly isochronous syllables ("scanning speech") and high variability in the durations of segments, are a hallmark of ataxic dysarthria, the speech disorder associated with cerebellar ataxia. However, the mechanisms underlying these temporal symptoms in ataxic dysarthria are unknown. One possibility is that speakers with ataxia also have an impaired ability to modulate timing in speech, including the process of detecting and correcting for errors that is critical to maintain movement accuracy. A robust body of literature has shown that neurobiologically healthy speakers continuously attend to sensory feedback to maintain speech accuracy, shown by adapting their speech production to counteract externally introduced perturbations of the auditory feedback they receive about their own speech (Houde & Jordan, 1998; Jones & Munhall, 2000; Purcell & Munhall, 2006).

Research in both the speech and non-speech domains has shown that people with ataxia do have impairments in sensorimotor adaptation (Criscimagna-Hemminger et al., 2010; Martin et al., 1996; Maschke et al., 2004; Morton & Bastian, 2006; Parrell et al., 2017; Statton et al., 2018). However, existing studies have focused on the spatial dimension of movement (e.g., visual perturbations of reaching angle or auditory perturbations of vowel formants), and there is currently no evidence regarding potential adaptation impairments in the temporal domain. Sensorimotor adaptation is perhaps even more vital for temporal control than for spatial control: it requires ~100-150 ms to see measurable compensatory responses to auditory feedback errors in the motor output (Rohde & Ernst, 2016) but this delay is potentially too long for the control of many speech segments (e.g., alveolar taps and unstressed schwa are typically sub-100 ms in duration).

Here, we examine potential impairments in temporal adaptation in speakers with ataxia relative to an age-matched group of neurobiologically healthy speakers. Specifically, we test the ability of both groups of speakers to adapt to an externally-introduced lengthening of the vowel /ɛ/ in "best", following recent work that has shown that neurobiologically healthy speakers exhibit robust temporal adaptation in vowels (Karlin et al., 2021; Oschkinat & Hoole, 2020). Critically, as it is impossible to react online to a lengthened vowel by shortening it, any shortening observed must be driven by adaptation in predictive or feedforward control of speech timing, rather than within-trial compensation for perceived errors.

Methods. 23 participants with ataxia and 36 age-matched neurobiologically healthy speakers have participated in the study out of an anticipated 40 per group. Data from 16 participants with ataxia and 16 age-matched neurobiologically healthy speakers has been analyzed to date and is presented here. The experiment had four phases, with a total of 90 trials: a 20-trial baseline phase with veridical feedback; a 20-trial ramp phase with incrementally increasing perturbation; a 30-trial hold phase at maximum perturbation; and a 20-trial washout phase with veridical feedback. On each trial, participants produced the target word "best". During the perturbation phases, the vowel /ɛ/ was lengthened (end of the segment was delayed), with a maximum lengthening of 60 ms (Figure 1A). The remainder of the word was played back at this delay, but with no additional perturbation to the duration of either /s/ or /t/. Perturbation was implemented using Audapter (Cai et al., 2010).

Data was automatically segmented by Audapter's OST function and then hand-corrected by the first author. Duration adaptation of /ɛ/ was measured as change from baseline at two points: 1) the last 10 trials of the hold phase, and 2) in the first 10 trials of washout; participant-specific baseline durations were taken as the mean of the last 10 trials of the baseline phase. Adaptation in this experiment would be reflected in shortening the vowel in hold and/or washout compared to baseline. Data was analyzed in R (R Core Team, 2019) using the lme4 package for linear mixed effects models (Bates et al., 2014). The full model included fixed effects of phase, group, and their interaction, and random intercepts by participant. Post-hoc tests were conducted with the emmeans package (Lenth, 2019). Estimated means are reported as change from baseline: negative values indicate shortening, and positive values indicate lengthening.

Results. Only neurobiologically healthy speakers showed adaptive shortening in /ɛ/, indicated by a significant improvement of model fit with the interaction between group and phase ($\chi^2(2) = 55.01$, $p < 0.0001$ compared to a model with group and phase alone); Figure 1B. Neurobiologically healthy speakers significantly shortened /ɛ/ in hold (-18.5 ± 3.8 ms) and in washout (-21.3 ± 3.8 ms, both $p < 0.0001$ compared to baseline); there was no significant difference between hold and washout ($p = 0.78$), indicating that these speakers maintained shortened productions through the first

10 trials of the washout phase. In contrast, speakers with ataxia did not change their /ɛ/ productions during hold (3.2 ± 3.8 ms) or washout (-3.8 ± 3.8 ms, both $p > 0.50$ compared to baseline), indicating that they were unable to adapt their productions. The neurobiologically healthy and ataxia groups are also significantly different from each other both in hold ($p = 0.003$) and washout ($p = 0.02$).

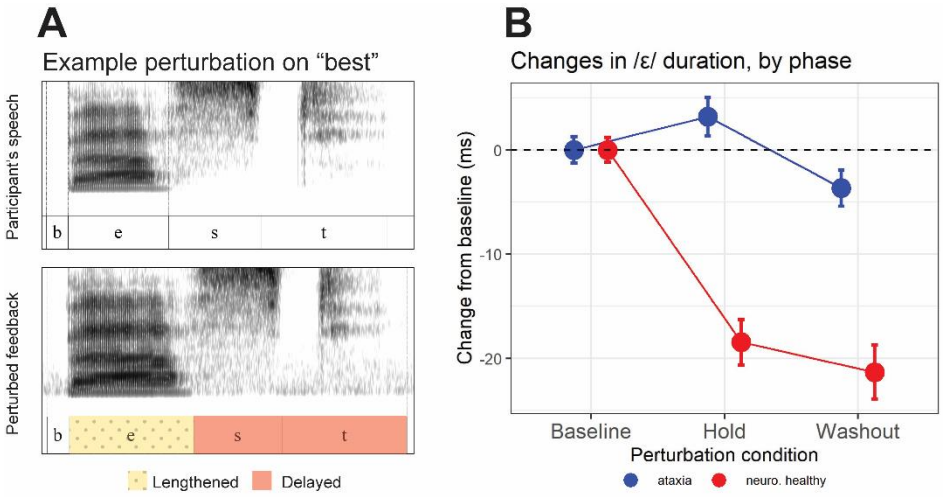


*Figure 1. **A**: Example showing maximal, 60-ms lengthening of the target vowel. **B**: Change in the duration of the target vowel /ɛ/ by group, normalized to the duration in the baseline phase.*

Discussion. The data indicates that, while age-matched neurobiologically healthy speakers showed robust adaptive shortening of the target vowel in response to the lengthened vowel duration in their auditory feedback, speakers with ataxia do not show any evidence of temporal adaptation. It is doubtful that this effect could result from a ceiling on movement speed: participants spoke at a self-selected, comfortable rate and it is unlikely that the participants with ataxia were speaking as quickly as possible at baseline. Although this result generally aligns with studies of spatial control of movement in both the speech and non-speech domains that have shown impaired sensorimotor adaptation in people with ataxia, these previous studies have largely shown *reduced* sensorimotor adaptation in the group with ataxia, rather than a full elimination of the response (Criscimagna-Hemminger et al., 2010; Morton & Bastian, 2006; Parrell et al., 2017). The complete lack of adaptation observed here suggests that the cerebellum may be especially critical for the adaptive control of temporal aspects of movement. Given the inadequacy of compensation for temporal control of speech movements, a total elimination of the ability to correct for temporal errors after the fact is one possible source of temporal deficits in ataxic dysarthria.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., & others. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R Package Version*, *1*(7), 1–23.

Cai, S., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2010). Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/and its pattern of generalization. *The Journal of the Acoustical Society of America*, *128*(4), 2033–2048.

Criscimagna-Hemminger, S. E., Bastian, A. J., & Shadmehr, R. (2010). Size of error affects cerebellar contributions to motor learning. *Journal of Neurophysiology*, *103*(4), 2275–2284.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354), 1213–1216.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America*, *108*(3), 1246–1251.

Karlin, R., Naber, C., & Parrell, B. (2021). Auditory Feedback Is Used for Adaptation and Compensation in Speech Timing. *Journal of Speech, Language, and Hearing Research*, *64*(9), 3361–3381.

Lenth, R. (2019). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans

Martin, T., Keating, J., Goodkin, H., Bastian, A., & Thach, W. (1996). Throwing while looking through prisms: I. Focal olivocerebellar lesions impair adaptation. *Brain*, *119*(4), 1183–1198.

Maschke, M., Gomez, C. M., Ebner, T. J., & Konczak, J. (2004). Hereditary cerebellar ataxia progressively impairs force adaptation during goal-directed arm movements. *Journal of Neurophysiology*, *91*(1), 230–238.

Morton, S. M., & Bastian, A. J. (2006). Cerebellar contributions to locomotor adaptations during splitbelt treadmill walking. *Journal of Neuroscience*, *26*(36), 9107–9116.

Oschkinat, M., & Hoole, P. (2020). Compensation to real-time temporal auditory feedback perturbation depends on syllable position. *The Journal of the Acoustical Society of America*, *148*(3), 1478–1495.

Parrell, B., Agnew, Z., Nagarajan, S., Houde, J., & Ivry, R. B. (2017). Impaired feedforward control and enhanced feedback control of speech in patients with cerebellar degeneration. *Journal of Neuroscience*, *37*(38), 9249–9258.

Purcell, D. W., & Munhall, K. G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America*, *120*(2), 966–977.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Statton, M. A., Vazquez, A., Morton, S. M., Vasudevan, E. V., & Bastian, A. J. (2018). Making sense of cerebellar contributions to perceptual and motor adaptation. *The Cerebellum*, *17*(2), 111–121.

Lingual motor learning based on real-time visual feedback in individuals with and without Parkinson's disease

Teja Rebernik[1,2*], Mark Tiede[3], Thomas B. Tienkamp[1], Defne Abur[1], Jidde Jacobi[1], Martijn Wieling[1]

[1] *University of Groningen*
[2] *Vrije Universiteit Brussel*
[3] *Yale University*
[*] t.rebernik@rug.nl

Introduction. Parkinson's disease (PD) is a neurodegenerative disease that affects motor movement. However, despite the motor difficulties that individuals with PD (IwPD) face, not all facets of motor *learning* seem to be affected. Implicit learning, which refers to unintentionally learning complex information (Seger 1994), seems to be mostly preserved in IwPD, with only motor skill retention (as opposed to motor skill acquisition) potentially impacted (Nieuwboer et al. 2009; Marinelli, Quartarone, et al. 2017). This seems further corroborated by IwPD's performance on sensory adaptation tasks, as IwPD on medication respond similarly to control participants in perturbation tasks that require integration of visual (Venkatakrishnan et al. 2011; Marinelli, Crupi, et al. 2009), tactile (Smiley-Olen et al. 2002) and auditory (Abur et al. 2021) feedback. The goal of our study was to assess the ability of IwPD to acquire a new lingual motor skill in an implicit speech-related task when real-time visual feedback is provided. Based on prior studies on implicit motor learning in IwPD, we expect no training effect difference between the two groups.

Methods. The lingual motor learning task formed part of a larger study, which has been approved by our Institutional Medical Ethics Review Board. A total of 43 Dutch native speakers completed the task, including 21 individuals with Parkinson's disease (IwPD; 11 male, 10 female; mean age 68.5 $\pm$ 8.7 years) and 22 control speakers (CS; 12 male, 10 female; mean age 67.7 $\pm$ 7.2 years). All IwPD completed the MDS-UPDRS assessment of symptom severity (Goetz et al. 2008), with scores on Part 3 ("motor symptom severity") ranging from 11-83 points. They did the task while ON levodopa.

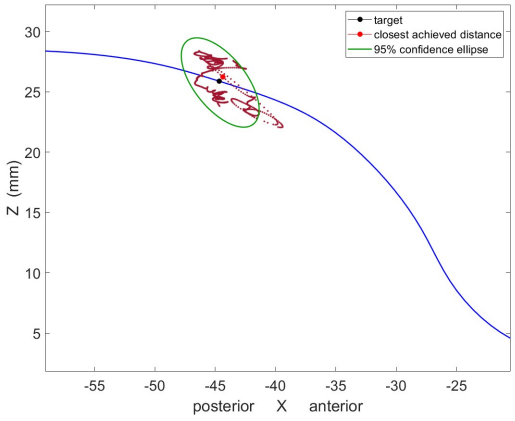


Figure 1: *An example view of a trial in analysis. The black dot on the palate (blue line) represents the middle of the posterior target, while the dark red lines represent all tongue movement near the target captured with the tongue tip sensor. The red dot is the minimum achieved distance to target, used for analysis.*

The lingual motor learning task was a real-time visual feedback task that required the participants to reach two palatal targets. We first placed NDI VOX electromagnetic articulography sensors (NDI VOX-EMA; Rebernik et al. 2021) on the mastoids and nasion as reference sensors. Afterwards, we collected biteplane and palate trace recordings, ensuring that each participant would be seeing their own palate. We then placed the tongue tip sensor, one centimetre from the anatomical tongue tip. During the experiment, the participants saw their palate with two targets superimposed. The first

target was an anterior target, placed at around 20% of the palate length, while the second target was a posterior target, placed at around 70% of the palate length. The anterior target constituted a familiar motor goal, as native speakers of Dutch place their tongue tip on the alveolar ridge during the production of alveolar consonants. The posterior target constituted an unfamiliar motor goal, as there is no sound in Dutch that would require speakers to form a constriction with their tongue tip further back on the palate. The participants were instructed to reach either the anterior or posterior target while receiving real-time head-corrected visual feedback of their tongue tip movement. The experimental task took around 10 minutes in total. There were 18 trials in the pre- and post-training condition, where participants had to tap the target only once, and 12 trials in the training condition, where participants had to tap the same target five times in a row.

For our analysis, we used a custom MATLAB script to extract the minimum distance to target that the participant reached in every trial ("minimum distance", lower values reflecting higher task performance accuracy) and the time point at which that occurred ("time-to-target"). See Figure 1 for an example trial. We expected a lower minimum distance and a faster response (taking less time) after training. In addition, we expected the posterior target to be more difficult to reach than the anterior target. Finally, we did not expect training effect differences between the groups. We built linear mixed-effects models using the *lme4* package in R version 4.3.1. Our hypothesis-testing model included z-transformed *time-to-target* and *minimum distance* values as the dependent variable (distinguished by the variable 'type'), and target (posterior vs. anterior), as well as a two-way interaction between test (pre-training vs. post-training) and group (PD vs. CS) as the fixed effects. The optimal random-effects structure included a by-participant random intercept, and type, test, and target as by-participant random slopes. The significance threshold alpha was set at 0.05. We conducted an additional exploratory analysis to determine the best model (via model comparison, using the *anova* function).

Results. In our hypothesis-testing model, target was significant ($\beta = 0.11, p = 0.04$) with the posterior target being more difficult. The interaction between test and group was not significant ($p = 0.33$). Separately, group showed a significant effect ($\beta = 0.18, p = 0.01$) with IwPD performing worse than controls. There was no significant training effect ($\beta = -0.06, p = 0.20$). The best exploratory model, however, revealed a training effect, but only for time-to-target ($\beta = -0.13, p = 0.01$) and not for the accuracy. This effect did not differ between the two groups ($p = 0.32$).

Discussion. The results indicate that there are no differences between IwPD and controls in how they use real-time visual feedback for lingual motor learning, further affirming that implicit learning skills are preserved in IwPD. However, there does seem to be a difference between the two groups overall, as IwPD took longer to hit the targets and did so less accurately compared to controls. This is also in line with studies that show potential reduced tactile acuity of the tongue tip in IwPD (e.g., Chen and Watson 2017). Interestingly, while there was a beneficial training effect for both groups for one of the measures (time-to-target), there was no benefit of training on the accuracy for both groups. Of course, when the target is reached in less time, it is likely that this comes at the cost of lower accuracy. Given that there was no reduction in accuracy after testing, but the targets were reached in less time, we may conclude that the (short) training was effective.

References.

Abur, D., A. Subaciute, A. Daliri, R. A. Lester-Smith, A. A. Lupiani, D. Cilento, N. M. Enos, H. R. Weerathunge, M. C. Tardif, and C. E. Stepp (2021). "Feedback and Feedforward Auditory-Motor Processes for Voice and Articulation in Parkinson's Disease". In: *JSLHR*.

Chen, Y.-W. and P. J. Watson (2017). "Speech production and sensory impairment in mild Parkinson's disease". In: *JASA*.

Goetz, C. G, B. C Tilley, S. R Shaftman, G. T Stebins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B Stern, R. Dodel, et al. (2008). "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results". In: *Movement Disorders* 23.15, pp. 2129–2170.

Marinelli, L., D. Crupi, A. Di Rocco, M. Bove, D. Eidelberg, G. Abbruzzese, and M. F. Ghilardi (2009). "Learning and consolidation of visuo-motor adaptation in Parkinson's disease". In: *Parkinsonism & Related Disorders*.

Marinelli, L., A. Quartarone, M. Hallett, G. Frazzitta, and F. G. Ghilardi (2017). "The many facets of motor learning and their relevance for Parkinson's disease". In: *Clinical Neurophysiology*.

Nieuwboer, A., L. Rochester, L. Müncks, and S. P. Swinnen (2009). "Motor learning in Parkinson's disease: limitations and potential for rehabilitation". In: *Motor learning in Parkinson's disease: limitations and potential for rehabilitation* 15.

Rebernik, T., J. Jacobi, M. Tiede, and M. Wieling (2021). "Accuracy assessment of two electromagnetic articulographs: Northern Digital Inc. WAVE and Northern Digital Inc. VOX". In: *JSLHR*.

Seger, C. A. (1994). "Implicit learning". In: *Psychological Bulletin*.

Smiley-Olen, A. N., H.-Y. K. Cheng, D. L. Latt, and M. S. Redfern (2002). "Adaptation of vibration-induced postural sway in individuals with Parkinson's disease". In: *Gait & Posture*.

Venkatakrishnan, A., J. P. Banquet, Y. Burnod, and J. L. Contreras-Vidal (2011). "Parkinson's disease differentially affects adaptation to gradual as compared to sudden visuomotor distortions". In: *Human Movement Science*.

Oral session 3

Production/Perception

12:00 am– 01:00 pm

	Title	Authors
12:00 - 12:20 am	Active inference and speech motor control: A review and theory	Abbie Bradshaw (University of Cambridge)*; Clare Press (University College London); Matt Davis (University of Cambridge)
12:20 - 12:40 am	Cross-linguistic interference and the perception-production relationship in L3 sound pronunciation	Yevgeniy Melguy (Basque Center on Cognition, Brain and Language)*; Clara Martin (Basque Center on Cognition, Brain and Language); Arthur Samuel (Basque Center on Cognition, Brain and Language)
12:40 am - 1:00 pm	Does auditory verbal aphantasia affect rhyme judgment?	Téo Pesci (Laboratoire de Psychologie et NeuroCognition)*; Jérémie Josse (Laboratoire de Psychologie et NeuroCognition); Alan Chauvin (Laboratoire de Psychologie et NeuroCognition); Romain Grandchamp (Laboratoire de Psychologie et NeuroCognition); Sharon Peperkamp (Laboratoire de Sciences Cognitives et Psycholinguistique); Hélène Loevenbruck (Laboratoire de Psychologie et NeuroCognition

Active inference and speech motor control: A review and theory

Abigail R. Bradshaw[1], Clare Press[2,3], Matt Davis[1]

[1]MRC Cognition and Brain Sciences Unit, University of Cambridge
[2] Experimental Psychology, University College London
[3]Wellcome Centre for Human Neuroimaging, University College London
abbie.bradshaw@mrc-cbu.cam.ac.uk

Introduction. Models of speech motor control place great emphasis on the prediction of sensory feedback, with sensory prediction errors being used to inform and modify movements (Guenther, 2016; Parrell & Houde, 2019). This mirrors claims in 'active inference' accounts; domain-general theories of brain functioning which reconceptualize the nature of the perception-action interface in terms of a common process of minimization of prediction errors (Adams et al., 2013; Friston et al., 2010). Such accounts have been extensively applied to the control of manual action and visual sensory feedback (e.g. Friston, 2011; Limanowski & Friston, 2020); however, they have received relatively little attention in speech motor control. We present here the first detailed application of an active inference framework to speech motor control, bridging the gap between these two literatures and suggesting new avenues for future research.

Methods. Our review first compares the architecture of active inference models to existing computational models of speech motor control; namely, the Directions Into Velocities of Articulators (DIVA) model (Tourville & Guenther, 2011) and the State Feedback Control (SFC) account (Houde & Nagarajan, 2011). We highlight similarities between these models, as well as areas of difference that might yield hypotheses for adjudicating between them. We then illustrate how active inference would account for compensation and speech motor adaptation following perturbations of auditory feedback.

Results. A comparison of the models found that active inference has much in common with both DIVA and SFC, with all three accounts sharing the central tenet that sensory prediction errors can be minimized both through action and through prediction updating; that is, an updating of stored internal models which specify the mappings between auditory outcomes and motor (or in the case of active inference, proprioceptive) targets. Active inference refers to such mappings as a 'generative model'; an internally constructed model of the outside world which generates sensory predictions. Active inference however differs from these models in several ways; e.g. the replacement of motor commands with proprioceptive predictions (and thus the exclusive reliance on a 'feedback' mode of motor control which contrasts with DIVA), the use of a shared set of predictions across both perception (of the self and others) and action (suggesting a unified account of speech perception and production), and the conceptualization of predictions as probability distributions rather than discrete targets or regions in SFC and DIVA. We provide a detailed description of an active inference account of compensation and adaptation to random versus sustained perturbations of speech auditory feedback. This demonstrates the importance of considering the role of predictions and sensory feedback in the proprioceptive domain, which form an integral part of the generative model. Specifically, we demonstrate how updating of proprioceptive predictions enables the translation of auditory prediction errors into changes to ongoing movement. Crucially, we also demonstrate how this updating is affected by changes in the precision or uncertainty of sensory predictions and sensory feedback. When perturbations are sustained, the precision of auditory feedback is increased, resulting in greater updating of predictions (and thus larger, long-lasting adaptation); conversely, when perturbations vary randomly from trial to trial, the precision of auditory feedback is decreased, resulting in less updating of predictions (and thus smaller, shorter-lasting compensation). In this way, our active inference account reduces the distinction between compensation and adaptation processes from a qualitative to a quantitative one, based on cross-trial inferences concerning the stability and volatility of auditory feedback during speech.

Discussion. We present here a preliminary demonstration of how active inference can be applied to speech motor control. In so doing, we highlight several emerging hypotheses and areas of interest for future research, to pave the way for further development and more detailed simulation of active inference accounts. In particular, we highlight the neglected role of proprioception in speech motor learning, and the need to consider the role of multimodal integration across auditory and proprioceptive feedback during speech. Furthermore, active inference accounts of speech offer the potential for reconceptualising the distinction between perception and action, placing perception and control of the self-voice and perception of other voices within a shared framework; this is likely to yield fresh insights into the interaction between the two, e.g. in phenomena such as phonetic convergence (whereby the voices of two interlocutors tend to acoustically converge to one another) (Pardo, 2006).

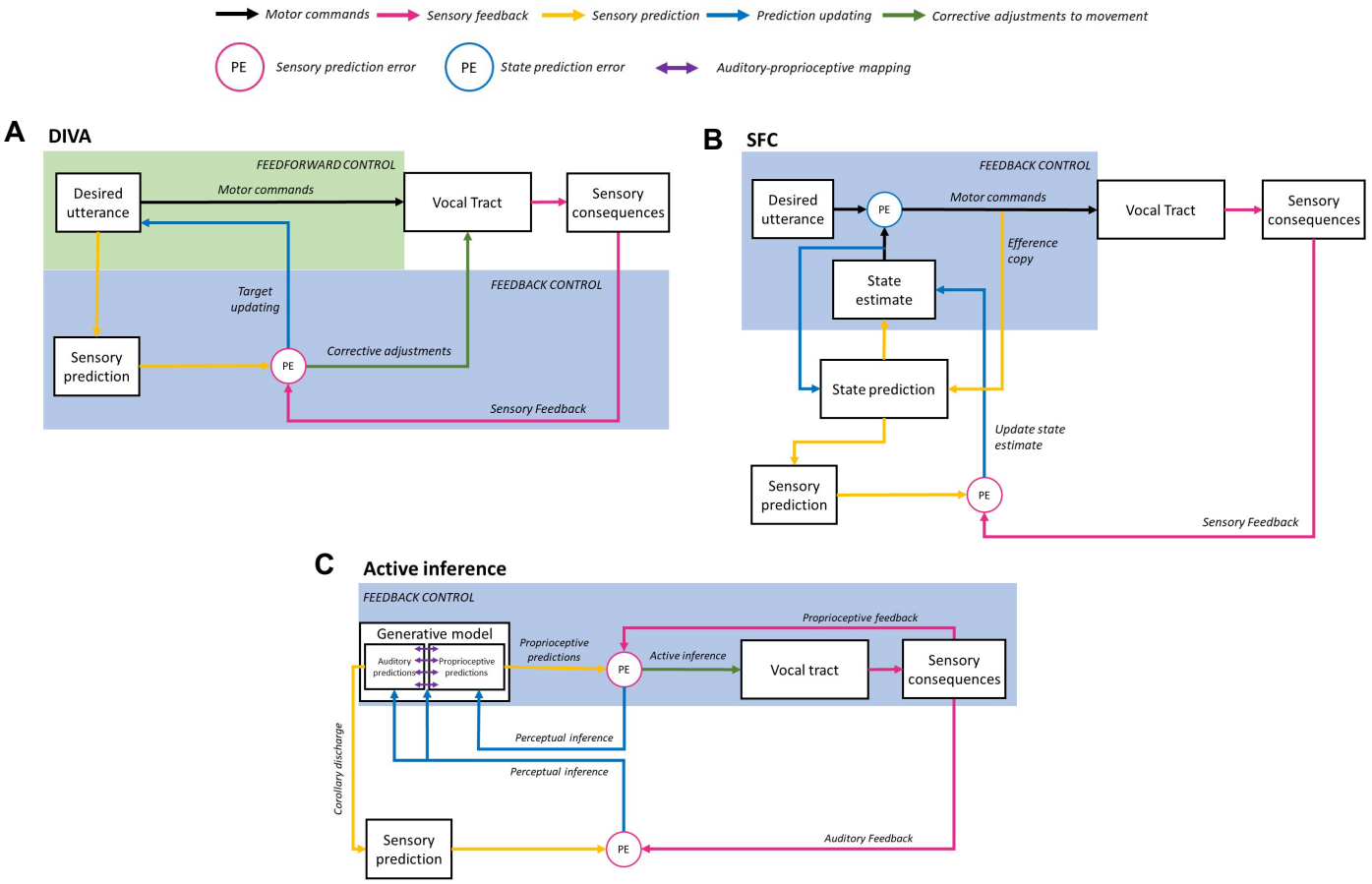


Figure 1: *Model architecture of (A) DIVA, (B) SFC and (C) active inference accounts of speech motor control.*

References

Adams, R. A., Shipp, S., & Friston, K. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, *218*(3), 611–643. https://doi.org/10.1007/s00429-012-0475-5

Friston, K. (2011). What Is Optimal about Motor Control? *Neuron*, *72*(3), 488–498. https://doi.org/10.1016/j.neuron.2011.10.018

Friston, K., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, *102*(3), 227–260. https://doi.org/10.1007/s00422-010-0364-z

Guenther, F. H. (2016). *Neural Control of Speech*. The MIT Press.

Houde, J., & Nagarajan, S. (2011). Speech Production as State Feedback Control. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00082

Limanowski, J., & Friston, K. (2020). Active inference under visuo-proprioceptive conflict: Simulation and empirical results. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-61097-w

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*(4), 2382–2393. https://doi.org/10.1121/1.2178720

Parrell, B., & Houde, J. F. (2019). Modeling the Role of Sensory Feedback in Speech Motor Control and Learning. *Journal of Speech Language and Hearing Research*, *62*(8, S, SI), 2963–2985. https://doi.org/10.1044/2019_JSLHR-S-CSMC7-18-0127

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952–981. https://doi.org/10.1080/01690960903498424

Cross-linguistic interference and the perception-production relationship in L3 sound pronunciation

Yevgeniy Melguy[1], Clara Martin[1,2], Arthur Samuel[1,2,3]

[1] Basque Center on Cognition, Brain, and Language
[2] Ikerbasque, Basque Foundation for Science
[3] Stony Brook University

y.melguy@bcbl.eu, c.martin@bcbl.eu, a.samuel@bcbl.eu

Introduction. Learners of a second language (L2) will often fail to achieve a native-like pronunciation, due to the influence of their L1 phonology. For instance, the Speech Learning Model or SLM (Flege 1995) claims that if a speaker perceives an L1-L2 sound pair as similar (e.g., /d/ in Spanish *dolor* "pain" vs. English *dollar*), learning is unlikely to occur: speakers will produce the L2 sound by substituting the closest L1 equivalent. The current study builds on this basic paradigm in several ways. First, while SLM was originally formulated to explain pronunciation in an L2, we extend it to L3 acquisition: what is the source of "accent" or cross-linguistic influence when multiple phonologies already exist in the learner's mind? Second, we address how individual differences may result in the success or failure of native-like L3 sound acquisition. SLM postulates that with sufficient experience, some learners will learn to perceptually distinguish between similar L1-L2 sounds and eventually form a new production category for the novel L2 sound. However, it does not specify the factors that will lead some learners to succeed, while others fail to ever approximate native-like production. Here, we investigate if individual differences among L3 learners (in global L3 proficiency as well as phonetic sensitivity) can account for differences in L3 sound acquisition.

Studies of L3 pronunciation have yielded mixed findings regarding the source of cross-linguistic influence. Some show transfer from the L1 (Llama & López-Morelos 2016), others from the L2 (Llama et al. 2010), and still others from both L1 and L2 (Sypiańska 2016). Interestingly, there is also evidence that the source of transfer may change over time, with transfer initially from the L2 but shifting to the L1 for more advanced speakers (Wrembel 2010). In addition to these possible group-level patterns, we should also see differences between individuals in the accuracy of their L3 production. Listeners have been shown to differ in their baseline perceptual sensitivity to subphonemic (within-category) differences (Kapnoula et al. 2017, Apfelbaum et al. 2022). While SLM would predict that individuals with higher phonetic sensitivity should be more likely to perceive (and consequently, to produce) differences between native vs. non-native sound pairs, this question has not been tested in the context of L3 acquisition. However, preliminary results (Kapnoula & Samuel 2021) suggest that listeners' perceptual sensitivity in the L1 predicts their overall L2 proficiency, suggesting that a similar relationship could hold for L3 segmental acquisition.

Methods. To test these questions, we examine sibilant fricatives (see **Figure 1**) in L1 Spanish - L2 Basque - L3 English speakers. While English has a two-way sibilant contrast (Collins & Mees 2003), Basque has a typologically rare 3-way contrast (Hualde et al. 2010). The sound missing from the English inventory (written <s> in Basque) is typically described in the literature as apical (involving the tongue tip) (Hualde et al. 2010), and is the same sound as the <s> in Castilian Spanish (Martínez Celdrán et al. 2003). Crucially, it differs from the English <s>: the latter is laminal (involving the tongue blade), sharing place of articulation with (unvoiced) Basque <z>. This study (in-progress) examines the perception and production of these sounds by 80 early Spanish-Basque bilingual speakers with a wide range of proficiency in L3 English, living in Donostia - San Sebastián, Spain. We first assess speakers' productions by obtaining acoustic measures (spectral center of gravity) of these sounds in each language via a picture-naming task, and then via imitation of a native English speaker in a shadowing task. We then assess perception via categorization of a phonetic continuum between each sibilant contrast (e.g., ship → sip), and then extract the slope for each listeners' categorization function to obtain a measure of sensitivity (Kapnoula et al. 2017). Thus, for each participant we will have several measures: (1) objective and subjective measures of overall proficiency in all three languages (these are available for all participants tested at the Basque Center on Cognition, Brain, and Language), and (2) a measure of sensitivity to subphonemic (within-category) phonetic differences in the L2 and L3. The accuracy of L3 production will be assessed by computing a difference score for each participant. In the picture naming task, this will be done by taking the difference in center of gravity for a participant's production of Basque <s> vs. their English <s>. In the imitation task, this will be the difference between the native English speaker's <s> productions and participants' shadowed productions. The perception-production relationship will thus be assessed by testing if overall L3 proficiency and/or perceptual sensitivity are significant predictors of L3 pronunciation accuracy.

Predicted results. This design allows us to test several possible sources of "accent" in L3 English. Listeners could transfer their L1 Spanish <s>, yielding an articulatory mismatch with English <s> (but an orthographic match), or they could transfer Basque <z>, a better articulatory-phonetic fit (but an orthographic mismatch). We also predict that listeners with higher perceptual sensitivity will be more likely to perceive (and produce) differences between similar

L1/L2 vs. L3 sounds, in line with the classic SLM assumption about the perception-production relationship. Also in line with the SLM prediction that more experienced listeners are more likely to acquire distinct production categories for non-native sounds, we predict that higher *global* L3 proficiency (i.e., lexical/grammatical knowledge) will predict higher L3 production accuracy. Finally, based on previous findings by Wrembel (2010), who found that the source of cross-linguistic influence in the L3 changes over time (shifting from L2 to L1 as learners become more proficient), we may find similar asymmetries in our data, with lower-proficiency learners showing transfer from L2 Basque <z>, but higher-proficiency L3 English speakers showing more transfer from L1 Spanish <s>. Thus, perhaps counter-intuitively, higher-proficiency L3 English speakers may actually show reduced production accuracy for this sound, since Spanish <s> is a poorer articulatory-phonetic fit to English <s> than Basque <z> is.

Discussion. This test case provides a window into how multiple existing sound systems may interact during L3 acquisition, and how such interaction is modulated by individual differences. Because the (Castilian) Spanish sibilant inventory is so limited, while Basque has an unusually rich set of sibilant contrasts, patterns of cross-linguistic influence in our speakers' L3 English will provide a valuable contribution to the relatively limited set of studies on L3 phonological acquisition. Moreover, by assessing the relationship between perceptual sensitivity and pronunciation accuracy in an L3, we hope to test whether core SLM assumptions about the perception-production relationship hold in the context of L3 sound acquisition.

Articulation	Sound	Spanish	Basque	English
palato-alveolar	[ʃ]	✘	✔ <x>	✔ <sh>
apical alveolar	[s̺]	✔ <s>	✔ <s>	✘
laminal alveolar	[s̻]	✘	✔ <z>	✔ <s>

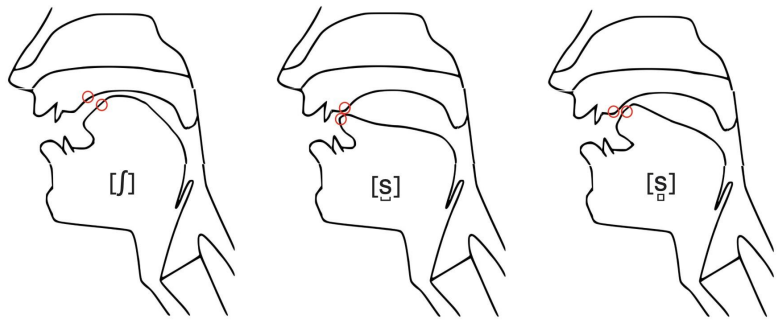


Figure 1. Mid-sagittal diagrams illustrating sibilant fricative place of articulation across three languages. Orthographic symbols for each sound are enclosed in angle brackets < >, and IPA symbols in square brackets [].

References

Apfelbaum, K. S., Kutlu, E., McMurray, B., & Kapnoula, E. C. (2022). Don't force it! Gradient speech categorization calls for continuous categorization tasks. *The Journal of the Acoustical Society of America*, *152*(6), 3728–3745.

Collins, B. D., & Mees, I. (2003). The phonetics of English and Dutch. In *The Phonetics of English and Dutch*. Brill.

Flege, J. E. (1995). Second Language Speech Learning Theory, Findings, and Problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. York Press.

Hualde, J. I., Lujanbio, O., & Zubiri, J. J. (2010). Goizueta Basque. *Journal of the International Phonetic Association*, *40*(1), 113–127.

Kapnoula, E. C., & Samuel, A. G. (2021). *Does sensitivity to acoustic variation within an L1 phoneme category help L2 learning?* in *Proceedings of the 62nd Annual Meeting of the Psychonomic Society*.

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(9), 1594–1611.

Llama, R., Cardoso, W., & Collins, L. (2010). The influence of language distance and language status on the acquisition of L3 phonology. *International Journal of Multilingualism*, *7*(1), 39–57.

Llama, R., & López-Morelos, L. P. (2016). VOT production by Spanish heritage speakers in a trilingual context. *International Journal of Multilingualism*, *13*(4), 444–458.

Martínez-Celdrán, E., Fernández-Planas, A. Ma., & Carrera-Sabaté, J. (2003). Castilian Spanish. *Journal of the International Phonetic Association*, *33*(2), 255–259.

Sypiańska, J. (2016). L1 vowels of multilinguals: The applicability of SLM in multilingualism. *Research in Language*, *14*(1), 79–94.

Wrembel, M. (2010). L2-accented speech in L3 production. *International Journal of Multilingualism*, *7*(1), 75–90.

Does auditory verbal aphantasia affect rhyme judgment?

Téo Pesci[1], Jérémie Josse[1], Alan Chauvin[1], Romain Grandchamp[1], Sharon Peperkamp[2], Hélène Lœvenbruck[1]

[1]Laboratoire de Psychologie et NeuroCognition, Université Grenoble Alpes
[2] Laboratoire de Sciences Cognitives et Psycholinguistique, École normale supérieure

Téo Pesci teo.pesci@univ-grenoble-alpes.fr

Pitch: can we have inner speech without inner auditory and motor sensation?

Introduction. When you imagine reading this title, you are perhaps able to hear your own voice and experience auditory sensations (e.g., intensity, intonation, etc.) although your auditory system is not being triggered by any external stimuli. This ability is called auditory verbal imagery. This kind of imagery can sometimes accompany endophasia. Endophasia (also called inner speech) refers to the internal production of language, without articulation or sound. Endophasia may be accompanied by auditory percepts (Langland-Hassan, 2018.; Sato et al., 2004), but this is not systematically the case: some people have aphantasia. Aphantasia refers to the lack of voluntary mental imagery (Zeman et al., 2015). Using questionnaires, Dawes et al. (2020) have shown the multisensory nature of aphantasia. Some individuals may have aphantasia in all sensory modalities ("profound" aphantasia), and others only in one sensory modality, for example a purely auditory aphantasia.

In this work, we focus on auditory verbal aphantasia (AVA) which refers to the inability to deliberately produce internal speech sounds. People with AVA report experiencing endophasia without any auditory sensation. This could suggest that individuals with AVA might not have access to the endophasia production stages that give rise to inner sound sensation. In the ConDialInt model of inner speech production (Grandchamp et al., 2019), it is assumed that inner speech shares processes with overt speech, in particular phonetic encoding (i.e., the transformation of amodal phonological representations into inner auditory sensations, through the mechanism of efference copy). We hypothesize that when speaking internally, people with AVA use early amodal phonological representations, without recourse to the auditory transformation process, which would explain the lack of sensory correlate.

To test this hypothesis, we compared the performance of a group of people with AVA to a control group performing a picture-prompted silent rhyme judgement task which is assumed to require access to phonetic encoding and for which the use of auditory correlates should improve performance (e.g., Rudner et al., 2019). This task consists in judging whether two words (illustrated by pictures but not written, nor orally presented) rhyme. The task was performed under two conditions: during articulatory suppression or during foot tapping. We used an articulatory suppression task to interfere with phonetic encoding (Gerwien et al., 2022; Wheeldon & Levelt, 1995). The following hypotheses were made: (1) if participants with AVA do not need to mentally generate auditory word forms and only use early available amodal phonological information, their judgments could be **faster than those of controls;** (2) if controls have access to multiple sensory information (auditory and articulatory), their responses could be **more accurate** than those of the group with AVA. We also expected to find an interaction between the group and the condition: controls should be slower and make more errors during articulatory suppression than during foot tapping compared with participants with AVA. Indeed, given that articulatory suppression interferes with the phonetic encoding stage and that we assume that participants with AVA do not have access to this stage in endophasia, they should not be impacted.

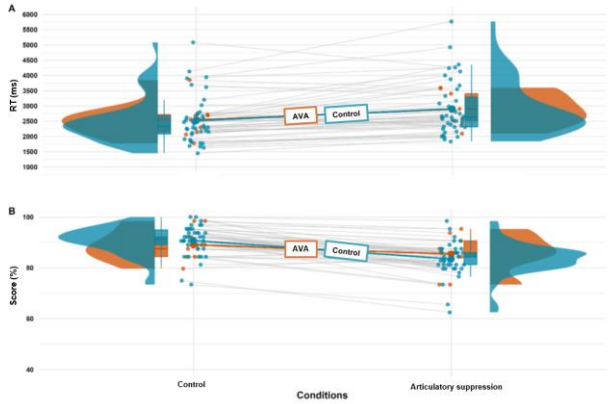
Methods. For each trial, participants had to judge whether or not the illustrated word pairs rhymed without saying them aloud. The task included two conditions: a control condition and an articulatory suppression condition. In both cases, a double task was performed. In the control condition, the double task consisted in judging a rhyme while tapping one's foot and passively listening to a voice whispering "patilon" over and over in headphones. In the articulatory suppression condition, participants had to whisper "patilon" continuously while judging whether the images rhymed or not. Comparison between the control condition and the articulatory suppression condition reveals the specific articulatory suppression effect on rhyme judgments. Each block was performed twice in a fixed order: control first then articulatory suppression.

Results. Correct response time were analyzed using linear mixed-effects models. For the accuracy score, we ran generalized linear binomial mixed-effects models.

Concerning response times, contrary to our expectation, results show that participants with auditory verbal aphantasia did not have significantly faster response times (M = 2.70 s, SD = 1.37) than control participants (M = 2.70 s, SD = 1.59, b = 0.01, 95% CI [-0.15; 0.18], ES = 0.08, t = 0.154, p = .878). On the other hand, as expected, there was a significant effect of articulatory suppression, with shorter response times in the tapping condition (M = 2.52 s, SD = 1.43) compared to the articulatory suppression condition (M = 2.90 s, SD = 1.66, b = 0.12, 95% CI [0.07; 0.18], ES = 0.02, t = 4.914, p < .001).

Moreover, the interaction between articulatory suppression effect and group was not significant; the suppression effect was similar between the two groups (*b* = -0.02, 95% *CI* [-0.11; 0.08], *ES* = 0.05, *t* = -0.313, *p* = .756).
Concerning scores, we observed no significant difference between AVA participants (*M* = 87.24%, *SD* = 33.38) and control participants (*M* = 87.13%, *SD* = 33.5, *log-OR*= 0.005, *CI* 95% -0.37; 0.38], *ES* = 0.19, *z* = 0.03, *p* = .9768). However, we observed a significant effect of articulatory suppression, with higher scores in the tapping condition (M = 90.50%, SD = 29.40) compared to the condition with articulatory suppression (*M* = 83.80%, *SD* = 36.80, *log-OR* = - 0.51, CI 95% [-0.70; -0.31], *ES* = 0.09, *z* = -5.124, p < .001). Finally, we observe no significant interaction effect between group and condition (*log-OR* = 0.34, 95% *CI* [-0.05; 0.73], *ES* = 0.20, *z* = 1.725, *p* = .0846).

Figure 1: *Response Times and Scores by Group and Condition*.



Discussion. Contrary to our predictions, participants with AVA were not significantly faster neither more accurate than the control group. Furthermore, both groups were equally sensitive to the effect of articulatory suppression. Articulatory suppression may be a more difficult task than foot tapping, as it requires more complex motor sequences than foot tapping, and involves breathing, phonation, and articulation coordination. The greater effect of this condition relative to tapping, could therefore simply be due to its greater difficulty. Given that AVA participants and controls are similarly affected by this task, it is possible that the use of amodal representations - available once phonological encoding has been completed - is in fact sufficient to judge whether two words rhyme or not, contrary to our initial prediction. In other words, phonetic encoding would not be necessary for this task. However, contrary to the recommendations of Nedergaard et al. (2023) we did not control the performance of the secondary task (i.e., articulatory suppression). It is therefore possible that the primary task (rhyme judgments) may also have influenced the secondary task. An examination of the recordings made during the run seems to support this hypothesis. Some of the images were less obvious than others.
When words were difficult to retrieve and pairs were difficult to judge, we noted that the repetition of the pseudoword "patilon" was slowed down, or even suspended in some cases. This remains to be further investigated.
In conclusion, our results seem to suggest that some people are able to use amodal phonological representations to perform some endophasia tasks. If the absence of auditory verbal imagery is established, this finding has important theoretical implications, notably by calling into question the auditory nature of phonology in inner speech (Langland-Hassan, 2018).

References

Dawes, A. J., Keogh, R., Andrillon, T., & Pearson, J. (2020). A cognitive profile of multi-sensory imagery, memory and dreaming in aphantasia. *Scientific Reports*, 10(1). https://doi.org/10.1038/s41598-020-65705-7

Gerwien, J., von Stutterheim, C., & Rummel, J. (2022). What is the interference in “verbal interference”? *Acta Psychologica*, 230, 103774. https://doi.org/10.1016/j.actpsy.2022.103774

Grandchamp, R., Rapin, L., Perrone-Bertolotti, M., Pichat, C., Haldin, C., Cousin, E., Lachaux, J.-P., Dohen, M., Perrier, P., Garnier, M., Baciu, M., & Lœvenbruck, H. (2019). The ConDialInt Model: Condensation, Dialogality, and Intentionality Dimensions of Inner Speech Within a Hierarchical Predictive Control Framework. *Frontiers in Psychology*, 10, 2019. https://doi.org/10.3389/fpsyg.2019.02019

Hubbard, T. L. (2010). Auditory imagery: Empirical findings. *Psychological Bulletin*, 302-329.

Langland-Hassan, P. (2018). From Introspection to Essence: The Auditory Nature of Inner Speech. In P. Langland-Hassan & A. Vicente (Éds.), Inner Speech: New Voices. Oxford: *Oxford University Press.*

Nedergaard, J. S. K., Wallentin, M., & Lupyan, G. (2023). Verbal interference paradigms: A systematic review investigating the role of language in cognition. *Psychonomic Bulletin & Review*, 30(2), 464-488. https://doi.org/10.3758/s13423-022-02144-7

Rudner, M., Danielsson, H., Lyxell, B., Lunner, T., & Rönnberg, J. (2019). Visual Rhyme Judgment in Adults With Mild-to-Severe Hearing Loss. *Frontiers in Psychology*, 10. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01149

Sato, M., Baciu, M., Lœvenbruck, H., Schwartz, J.-L., Cathiard, M.-A., Segebarth, C., & Abry, C. (2004). Multistable representation of speech forms : A functional MRI study of verbal transformations. *Neuroimage* 23, 1143–1151. doi: 10.1016/j. neuroimage.2004.07.055. Psychol. Rev, 10-1037.

Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives without imagery – Congenital aphantasia. *Cortex*, 73, 378-380. https://doi.org/10.1016/j.cortex.2015.05.019

Poster 1

3:00 – 5:00 pm

Paper	Title	Authors
131	Articulatory Dynamics of Lexical Stress in L2 English: A Case Study of Taiwanese Mandarin Speakers	Paul McGuire (National Tsing Hua University)*; Feng-fan Hsieh (National Tsing Hua University); yueh-chin chang (NTHU)
222	Towards a minimal dynamics for gestures: a law relating velocity and position	Michael Stern (Yale University)*; Jason A Shaw (Yale)
42	Online compensation of auditory feedback perturbations in speech: an optimal feedback control model of tongue biomechanics	Ny T Rakotomalala (Gipsa-lab)*; Pierre Baraduc (GIPSA-lab); Pascal Perrier (Gipsa-lab, Grenoble INP, Université Grenoble Alpes)
55	Assessing differences in articulatory-acoustic vowel space in Parkinson's disease phenotypes	Nikki Hoekzema (University of Groningen); Teja Rebernik (University of Groningen)*; Thomas B Tienkamp (University of Groningen); Sasha Chaboksavar (University of Groningen); Valentina Ciot (University of Groningen); Annetje Gleichman (University of Groningen); Roel Jonkers (University of Groningen); aude noiray (LPNC- UGA); Martijn Wieling (University of Groningen); Defne Abur (University of Groningen)
31	Effects of expectedness and clarity of speech auditory feedback on perception and motor control	Abbie Bradshaw (University of Cambridge)*; Clément Gaultier (University of Cambridge); Clare Press (University College London); Matt Davis (University of Cambridge)
223	Quantifying the Lombard Effect in Noisy Environments	Jackie Kim (Boston University)*; Alan Bush (Massachusetts General Hospital/Harvard Medical School); Matteo Vissani (Massachusetts General Hospital/Harvard Medical School); Mark Richardson (Massachusetts General Hospital); Frank H. Guenther (Boston University)
46	The Effect of Speaking Style on the Articulatory-Acoustic Vowel Space in Individuals with Tongue Cancer Before and After Surgical Treatment	Thomas B Tienkamp (University of Groningen)*; Teja Rebernik (University of Groningen); Raoul Buurke (University of Groningen); Katharina M. Polsterer (University of Groningen); Rob van Son (Netherlands Cancer Institute); Martijn Wieling (University of Groningen); Max J. H. Witjes (University Medical Center Groningen); Sebastiaan de Visscher (University Medical Center Groningen); Defne Abur (University of Groningen)
180	Aryepiglottic trilling in Mehweb: acoustics and variability	Ekaterina Shepel (National Research University Higher School of Economics); Alexandre Arkhipov (Universität Hamburg)*; Michael Daniel (Collegium de Lyon / Laboratoire Dynamique du Langage); Alexander Shiryaev (Independent researcher)
123	An investigation of syllable position /l/ allophony in L2 English learners using Word Error Rate as an index of phonetic proficiency	Anisia Popescu (Université Paris Saclay - LISN)*; Lori Lamel (CNRS LISN); Ioana Vasilescu (LIMSI); Laurence Y. Devillers (LISN-CNRS)
122	Acceleration peaks as representation of activation strength	Malin Svensson Lundmark (Lund University)*
186	Are long and short vowels articulatorily different?: Spatial and durational effects of vowel length	Sireemas Maspong (LMU Munich)*; Francesco Burroni (LMU Munich)
35	Enhancing lip contrasts between /u/ and /y/ in Taiwan Mandarin	Chenhao Chiu (National Taiwan University)*; Cheng-Hsiang Chang (National Taiwan University); Jian-zhi Huang (National Taiwan University); Po-Hsuan Huang (National Taiwan University)
247	Sub-visemic discrimination and the effect of visual resemblance on silent lip-reading	Maëva Michon (Praxiling)*
144	Acquiring tongue shape complexity in Scottish Gaelic consonants	Claire Nance (Lancaster University)*; Sam Kirkham (Lancaster University)
160	Effects of fundamental frequency and spectral manipulations on speech production under delayed auditory feedback	Yasufumi Uezu (Japan Advanced Institute of Science and Technology)*; Masato Akagi (Japan Advanced Institute of Science and Technology); Masashi Unoki (JAIST)
210	Perception and production are related in novice learners of Mandarin lexical tone	Jennifer Yang (New York University)*; Xi Chen (New York University); Joyce Chung (Boston University); Charles Chang (Boston University); Tara McAllister (New York University)
47	Auditory Vowel Discrimination in Middle Childhood Compared to Adulthood	Katharina M. Polsterer (University of Groningen)*; Nikki Hoekzema (University of Groningen); Xingfeng Yang (University of Groningen); Thomas B Tienkamp (University of Groningen); Teja Rebernik (University of Groningen); Hedwig Sekeres (University of Groningen); Reihaneh Amooie (University of Groningen); Raoul Buurke (University of Groningen); Wietse de Vries (University of Groningen); Liyang Wang (University of California, Berkeley); Frank Tsiwah (University of Groningen); Martijn Wieling (University of Groningen); Defne Abur (University of Groningen)
10	Speaking-induced Middle Ear Muscle Reflex (MEMR): suppression of auditory feedback during self-vocalization	Hayo Terband (Department of Communication Sciences and Disorders, University of Iowa)*; Caroline Cross (Department of Communication Sciences and Disorders, University of Iowa); Shawn Goodman (Department of Communication Sciences and Disorders, University of Iowa)
153	Analysing the vocal tract front-back relationships	Antoine Serrurier (Uniklinik RWTH Aachen)*

204	Acoustic cues to lexical stress in Bulgarian	Milena Milenova (University of Sofia)*
74	An exploration of pitch in Afro-Mexican Spanish	Gilly Marchini (University of Edinburgh)*
188 (Remote)	Mandarin Chinese tonal coarticulation in the production of learners with an atonal L1	Kornélia Juhász (HUN-REN Hungarian Research Centre for Linguistics)*; Huba Bartos (HUN-REN Hungarian Research Centre for Linguistics)
6	A constriction geometry analysis of place contrasts in Malayalam nasals	Alexei Kochetov (University of Toronto)*; Pierre Badin (GIPSA-lab, Grenoble)
128	Effect of varying rhythmic stimulations on fluency and production gestures of people who stutter	Maëva GARNIER (GIPSA-lab)*; Anneke W. Sas (University of Wisconsin-Madison); Victor Adlard (GIPSA-lab); Christophe Savariaux (GIPSA-lab)
211	Prenasalization in initial voiced stops in Zuberoan Basque	Ander Egurtzegi (CNRS-IKER)*; Iñigo Urrestarazu Porta (CNRS-IKER, UPPA, UPV/EHU); Andrea Garcia Covelo (IPS-LMU Munich, IKER-UMR5478, UPPA)
93 (Remote)	Vocal expression of emotions in patients with unilateral vocal fold paralysis	Caterina Petrone (LPL)*; Nicolas Audibert (Laboratoire de Phonétique et Phonologie); Ralph Haddad (LPL/Hôpital La Conception); Méline Robert (LPL); Marion Tracq (LPL); Alexia Matteï (LPL); Lalain Muriel (LPL)
23	Lingual ultrasound feedback in L2 pronunciation practice in classroom: a pilot study of French mid front-back vowel contrast	Claire Pillot-Loiseau (Sorbonne Nouvelle University)*; Hélène Gustin-Masset (Sorbonne Nouvelle University); Tanja Kocjančič Antolík (Charles University); Takeki Kamiyama (Université Paris 8 Vincennes - Saint-Denis, TransCrit, Saint Denis)
177	Pinocchio, a biomimetic mechatronic tested for producing in vitro articulated speech	Nathalie Henrich Bernardoni (CNRS)*; Julien Royer (GIPSA-lab); Mourad Thaidi (GIPSA-lab); Xavier Laval (GIPSA-lab); Sylvain Arnaud (GIPSA-lab); Lucie Bailly (3SR)
171	A Biomechanical Tongue Model of a Neanderthal	Maxime Callia (SCD, Sorbonne Université)*; Pablo A Alvarez (Inria); Pascal Perrier (Gipsa-lab, Grenoble INP, Université Grenoble Alpes); Yohan PAYAN (Univ. Grenoble Alpes); Amélie Vialet (Muséum National d'Histoire Naturelle)
94	Acquisition of articulatory dynamics in second language speech: Japanese speakers' production of English and Japanese liquids	Takayuki Nagamine (Lancaster University)*
147	A dynamical model of diachronic vowel change	Sam Kirkham (Lancaster University)*; Patrycja Strycharczuk (University of Manchester)
17	An experimental setup for capturing multimodal accommodation using dual electromagnetic articulography, audio, and video	Lena Pagel (IfL Phonetics - University of Cologne)*; Simon Roessig (University of York); Doris Mücke (IfL Phonetics - University of Cologne)
133	Auditory Targets for Sensory Feedback Control of Speech Change Over the Course of the Day	Frank H. Guenther (Boston University)*; Alexander Acosta (Boston University); Elaine Kearney (Queensland University of Technology)
73	From YIN to β-YIN: algorithm optimisation and performance analysis on auto-oscillating vocal folds replicas for normal and abnormal conditions	Raphaël Chetien (LEGI CNRS)*; Armonie Van Hirtum (CNRS); Xavier Pelorson (CNRS); Didier Demolin (LPP CNRS)
137	Differential effect of contrast between self-produced and other-produced vowels on corrective vowel production in child and adult speech	Melissa A Redford (University of Oregon)*; Carissa Diantoro (University of Oregon)
224	The contribution of voicing to coarticulatory nasalization in two varieties of English	Conceição Cunha (Institute for Phonetics, LMU Munich)*; Jonathan Harrington (Institute for Phonetics, LMU Munich); Philip A Hoole (Institute of Phonetics, Munich University)
40	On the relation between breathing and utterance length in vocally learning birds	Susanne Fuchs (zas)*; Lara S. Burchardt (ZAS); Franz Goller (University Münster & University Utah)
167	Attentional demand on speech processing: evidence from dual-task interference on vowel space and V-to-V anticipatory coarticulation according to task properties	Michaela Pernon (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS-Université Sorbonne Nouvelle)*; Dana D'Alessandro (University of Washington)
239	Examining Speech Perception of Non-Errored Pronunciations in Children with Speech Sound Disorders	Elaine R. Hitchcock (Montclair State University)*; Laura L. Koenig (Adelphi University, Haskins Labs)
189	Auditory feedback of speech: comparison between the aerial and the bone-conducted pathway	Raphaël Vancheri (GIPSA-lab); Corentin E. Vilain (GIPSA-LAB); Nathalie Henrich Bernardoni (CNRS); Pierre Baraduc (GIPSA-lab)*
62	The production of speech modes in motor speech disorders	Marion Bourqui (University of Geneva)*; Monica Lancheros (University of Geneva); Frédéric Assal (Geneva University Hospital); Marina Laganaro (University of Geneva)
237	Objective measures of fatigue and sleepiness based on acoustic analysis of the temporal organization of speech	Hani C Yehia (UFMG - Universidade Federal de Minas Gerais)*; Déborah Abrante (UFMG - Universidade Federal de Minas Gerais); Carla Vasconcelos (UFMG - Universidade Federal de Minas Gerais); TÃºlio Rodrigues (USP - Universidade de São Paulo); Maurílio Vieira (UFMG - Universidade Federal de Minas Gerais)
184	The Interplay between Acoustics and Syllable Articulation Organized by Mandible Movement	Donna M Erickson (Haskins labs)*; Plínio A Barbosa (UNICAMP); Gustavo Silveira (University of Campinas)
83	Speech Intelligibility Decreases with Degradation of Somatosensory Feedback via Topical Benzocaine Application	Elizabeth Casserly (Trinity College)*; Anna Barnes (Trinity College); Lauren Barrett (Trinity College)
145	Manner does not affect articulatory overlap in Spanish voiced-stop+lateral clusters	Mark Gibson (Universidad de Navarra)*; Stavroula Sotiropoulou (Universität Potsdam); Adamantios Gafos (Universität Potsdam)

Articulatory Dynamics of Lexical Stress in L2 English: A Case Study of Taiwanese Mandarin Speakers

Paul McGuire, Feng-fan Hsieh, Yueh-chin Chang

National Tsing Hua University, Taiwan

graemepaulmcguire@gmail.com, ffhsieh@mx.nthu.edu.tw, ycchang@mx.nthu.edu.tw

Introduction. Stressed syllables are said to 'hyper-articulated' (de Jong, 1995) or to "involve longer, larger and faster gestures than their unstressed counterparts" (Katsika & Tsai, 2021). This exploratory study, involving 10 participants, utilised electromagnetic articulography (EMA) to examine the articulatory patterns of lexical stress minimal pairs in L2 English, as produced by native speakers of Taiwanese Mandarin, a typical East Asian tone language without the stressed vs. non-stressed contrast. Articulatory trajectories of vowels were analysed using generalised additive mixed (GAMM) modelling. Syllable initial consonant gestures were analysed in terms of gestural duration, peak velocity and amplitude-normalised peak velocity (stiffness; see Roon et al., 2021).

Methods. Ten native speakers of Taiwanese Mandarin were recruited for this study. All participants were in their twenties and spoke only Mandarin in their daily life in Taiwan. This study focussed on three disyllabic minimal pairs which differ only in stress location (CONflict - conFLICT, PROject - proJECT, DIgest - diGEST). The target words were embedded in the carrier phrase "Please say ____ again" and read in randomised order from a screen in a soundproof room. Eight participants read each word ten times, while the remaining two read each word seven times. Participants also read a paragraph from an AI-generated short story which was used to assess their level of accentedness. Articulatory data were recorded using EMA (Carstens AG501) at a sampling rate of 2,000 Hz, later down-sampled to 250 Hz. Sensors were attached to the lips, tongue, and lower incisor (for tracking jaw movement), as well as to the right and left mastoid processes and upper incisor (to correct for head movement). The sensors relevant to this study are TT (tongue tip), TB (tongue body), TD (tongue dorsum) and JAW (lower incisor). Acoustic data were recorded simultaneously at 24 kHz.

Articulatory measurements were made in Matlab using Mview (Tiede, 2005). For the vowel analysis, vocalic portions were identified using acoustic data. Time-normalised and within-speaker z-scored sensor trajectories were compared using generalised additive mixed modelling (GAMM) in R (based on recommendations from Wieling, 2018). Gesture durations for the consonant analysis, specifically the hold phase (i.e., NOFFS – NONS), were identified using the findgest() algorithm in Mview, which identifies gestural landmarks based on a peak velocity threshold of 20%. Statistical testing was carried out using linear mixed effects modelling with the lme4 package (Bates et al. 2015) in R.

Results. The results of the vowel analysis are presented in Table 1. Asterisks denote significantly different articulator trajectories between the stressed and unstressed vowels that are continuous for a portion comprising at least 15% of the vocalic section. Anatomical directions—superior, inferior, anterior, and posterior—refer to the position of the articulator in the stressed syllable (i.e. 'CON') relative to its position in the syllable's unstressed counterpart (i.e. 'con'), during the portion where significant difference is observed.

Table 1: Vowel GAMM analysis results (x = front/back; z = high/low)

	TDz	TDx	TBz	TBx	TTz	TTx	JAWz	JAWx
CON	* inferior		* inferior		* inferior		* inferior	
FLICT	* superior	* anterior		* anterior				
DI	* superior		* inferior		* inferior	* posterior	* inferior	* posterior
GEST					* inferior		* inferior	
PRO	* inferior		* inferior			* posterior	* inferior	* posterior
JECT					* inferior		* inferior	

The results indicate that stressed vowels were most consistently associated with larger jaw displacement, with all stressed vowels other than 'FLICT' showing significantly more inferior jaw positions than their unstressed counterparts. Hyper-articulation of the lingual articulators was also observed in at least one dimension in every stressed syllable. Plots of the vocalic section of 'DI' revealed two portions of significant difference, inferior TBz for the low vowel at the starting point of the diphthong and superior TDz for the high vowel at the end. Additionally, in the syllable 'CON', TDz and TBz are seen to be to have a second portion of significant difference (in a superior position relative to that in 'con') towards the end of the vocalic section. This tendency of L1 Mandarin speakers to realise alveolar nasals[1] as velar nasals in the context of a back vowel is due to what Duanmu (2007) terms 'Rhyme Harmony' and is often seen in loanword adaptation (e.g. Hsieh et al., 2009).

As for the syllable initial consonants, data were aggregated and normalised in R. Linear mixed-effects models were constructed for each of the three measurements of interest - gesture duration, peak velocity, and stiffness, using the lmer() function. Among these three variables, only gesture duration demonstrated a statistically significant association. This association was positive, indicating that gesture duration is longer in stressed syllables. Stiffness and peak velocity did not show a significant relationship with stress. Further analysis indicated that speakers with heavier accents did not exhibit significant differences in gesture durations between stressed and unstressed conditions. For each participant, we calculated the difference in gestural durations for stressed versus unstressed consonants. A statistically significant correlation emerged between these differences and the participants' accentedness scores. It was observed that the greater the perceived native-sounding quality of the participants' English, the more pronounced the difference was in the durations of consonant gestures between stress and unstressed syllables.

Discussion. This study found that stressed vowels are associated with larger jaw displacement and are hyper-articulated to some degree in Taiwanese Mandarin-accented English. Additionally, this study suggests that L2 learners, whose L1 lacks a stress distinction, maybe be able to acquire articulatory timings associated with L1 stress production, such as gestural plateau durations. The GAMM approach revealed patterns which might have gone unnoticed in studies that compare articulatory measurements from a static position, such as the centre of a vowel. Interestingly, Kim's (2021) results suggest that the stressed syllables do not involve substantial supra-glottal hyper-articulation in L2 English by Standard Chinese speakers. This discrepancy could be attributed to several potential confounding factors: the contrast between spontaneous and laboratory speech, the difference between point-to-point comparison and trajectory analysis of EMA sensors, and variations across Mandarin dialects.

Analyses of C-V gesture timing and acoustics are currently underway, and the results will be presented at the conference.

References.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Green, P. (2009). Package 'lme4'. URL http://lme4. r-forge. r-project. org.

De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. The journal of the acoustical society of America, 97(1), 491-504.

Duanmu, S. (2007). The phonology of standard Chinese. OUP Oxford.

Hsieh, F. F., Kenstowicz, M., & Mou, X. (2009). Mandarin adaptations of coda nasals in English loanwords. Loan phonology, 131-154.

Katsika, A., & Tsai, K. (2021). The supralaryngeal articulation of stress and accent in Greek. Journal of Phonetics, 88, 101085.

Kim, B. (2021). Lexical Stress Realization in Mandarin Second Language Learners of English: An Acoustic and Articulatory Study. (Ph.D. dissertation). CUNY Graduate Center

Roon, K. D., Hoole, P., Zeroual, C., Du, S. H., & Gafos, A. I. (2021). Stiffness and articulatory overlap in Moroccan Arabic consonant clusters. Laboratory Phonology, 12. DOI: 810.5334/labphon.272.

Tiede, M. (2005). MVIEW: software for visualization and analysis of concurrently recorded movement data. New Haven, CT: Haskins Laboratories.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. Journal of Phonetics, 70, 86-116.

Wood, S. (2019). mgcv: mixed GAM computation vehicle with automatic smoothness estimation. R-package version 1.8–31.

[1] The production of the coda nasal in 'CON/con' varied between speakers, with a nasalised vowel being the most frequent realisation. Consequently, the entire rime was included in the analysis.

Towards a minimal dynamics for gestures: a law relating velocity and position

Michael C. Stern[1], Jason A. Shaw[1]

[1]Department of Linguistics, Yale University

michael.stern@yale.edu, jason.shaw@yale.edu

Introduction. In controlled human movement—including speech articulatory movement—peak velocity is robustly correlated with maximum spatial displacement (Ostry & Munhall, 1985). The farther an effector travels to reach its target, the faster it moves. In order to capture this empirical fact, dynamical models of articulatory movement, e.g., Task Dynamics (Saltzman & Munhall, 1989), encode a negative relationship between velocity and displacement, of the form **(1)** $\dot{\boldsymbol{x}}_{\boldsymbol{t}} = -\boldsymbol{\lambda}(\boldsymbol{x}_{\boldsymbol{t}} - \boldsymbol{x}_{\boldsymbol{0}})$, where x_t is the state at time t of a vocal tract variable (TV) like lip aperture (LA: the distance between the lips), x_0 is the target state of the TV (e.g., 0 for a /b/ or /m/), and λ is a control parameter modulating the relationship between velocity $\dot{x}_t$ and displacement $(x_t - x_0)$. The system in (1) succeeds in capturing the linear correlation between peak velocity and maximum displacement. However, it fails to capture another robust fact about TV trajectories. In particular, for any fixed value of the control parameter λ, model-simulated TV trajectories achieve peak velocity instantaneously; velocity then decreases monotonically as the TV approaches its target. In real TV trajectories, velocity curves are approximately symmetrical: peak velocity is achieved approximately halfway through the movement (Ostry et al., 1987). In the *damped spring* model of Task Dynamics, as in **(2)** $\boldsymbol{b}\dot{\boldsymbol{x}}_{\boldsymbol{t}} = -\boldsymbol{m}\ddot{\boldsymbol{x}}_{\boldsymbol{t}} - \boldsymbol{k}(\boldsymbol{x}_{\boldsymbol{t}} - \boldsymbol{x}_{\boldsymbol{0}})$, peak velocity is delayed because velocity $\dot{x}$ is negatively related to acceleration $\ddot{x}$. This empirical improvement is achieved via greater model complexity: (2) is a *second order* system, referencing acceleration in addition to velocity, with three control parameters b, m, and k, rather than the single parameter λ. Even in (2), however, velocity curves are unrealistically right-skewed, with peak velocity occurring earlier than halfway through the movement (Perrier et al., 1988). Thus, additional complexity has been proposed: e.g., a time-varying *activation* parameter (Byrd & Saltzman, 1998; Kröger et al., 1995), or a negative relationship between velocity and the *cube* of displacement (Sorensen & Gafos, 2016).

We take an empirical approach to understanding the relation between velocity and position. Rather than commit to the specific second order system in (2), we first ask: what is the *empirical* relationship between velocity and displacement over time? The answer to this question can guide further model building, which we pursue below.

Methods. Our data come from electromagnetic articulography (EMA) measurements of CV syllables [ma], [mi], [ba], and [bi] produced in real words in carrier phrases. Each target syllable was preceded by a vowel (if the target vowel was [a], the preceding vowel was [i], and vice versa) in order to ensure maximal vowel movement. 12 speakers of English and 12 speakers of Mandarin produced 128 tokens each. Consonant constriction movements were parsed from the LA signal. Movement onsets were marked as the timepoint at which velocity surpassed 20% of the maximum, and movement offsets were marked as the timepoint at which velocity fell below 20% of the maximum. Spatial targets x_0 were defined as the LA value at the point of minimum velocity after movement offset. For each consonant constriction movement, we calculated λ_t at each sample t as the negative ratio of velocity at t to displacement at t: $\frac{-\dot{x}_t}{(x_t - x_0)}$. By demarcating movements based on a threshold of 20% of maximum velocity, instead of, e.g., velocity zero-crossing, we exclude portions of the kinematics in which velocity or displacement are infinitesimal. This prevents λ_t from approaching 0 (infinitesimal velocity) or infinity (infinitesimal displacement). After data cleaning, 1963 tokens remained for analysis.

Results & Discussion. Qualitative examination of λ trajectories suggests that λ generally follows an *exponential growth* pattern, as exemplified in **Figure 1**. The ratio of velocity to displacement grows exponentially from movement onset to target achievement. This qualitative characterization is supported by the fact that $\log(\lambda)$ provides an excellent linear fit to the data, with a mean R^2 of 0.960 for English and 0.957 for Mandarin. Moreover, the slopes of the linear fits correlate strongly with both gesture duration and kinematic stiffness (peak velocity divided by displacement: Roon et al., 2021), as seen in **Figure 2**. The log-linear nature of λ trajectories suggests the following dynamics for gestures: TV trajectories are governed by the simple first order law in (1), but with time-varying λ, i.e., λ_t, as in **(3)** $\dot{\boldsymbol{x}}_{\boldsymbol{t}} = -\boldsymbol{\lambda}_{\boldsymbol{t}}(\boldsymbol{x}_{\boldsymbol{t}} - \boldsymbol{x}_{\boldsymbol{0}})$. The evolution of the parameter λ_t is governed by the exponential growth law **(4)** $\boldsymbol{\lambda}_{\boldsymbol{t}} = \boldsymbol{e}^{\boldsymbol{r}} * \boldsymbol{\lambda}_{\boldsymbol{t-1}}$, derived from the empirical discovery that $\log(\lambda_t) = \log(\lambda_{t-1}) + r$. The system defined in equations (3) and (4) has one control parameter r, which can be directly inferred from data and correlates strongly with linguistically relevant measures like duration and stiffness (**Figure 2**). Moreover, TV trajectories simulated from this model display more-or-less symmetrical velocity curves, as seen in **Figure 3**. Thus, a second order system like the damped spring model (with additional complexity like ramped activation or cubic non-linearity) is not strictly necessary to capture symmetry in the velocity profile. Future work will extend our model to other tract variables besides lip aperture and other types of gestures like consonant release gestures and vowel gestures.

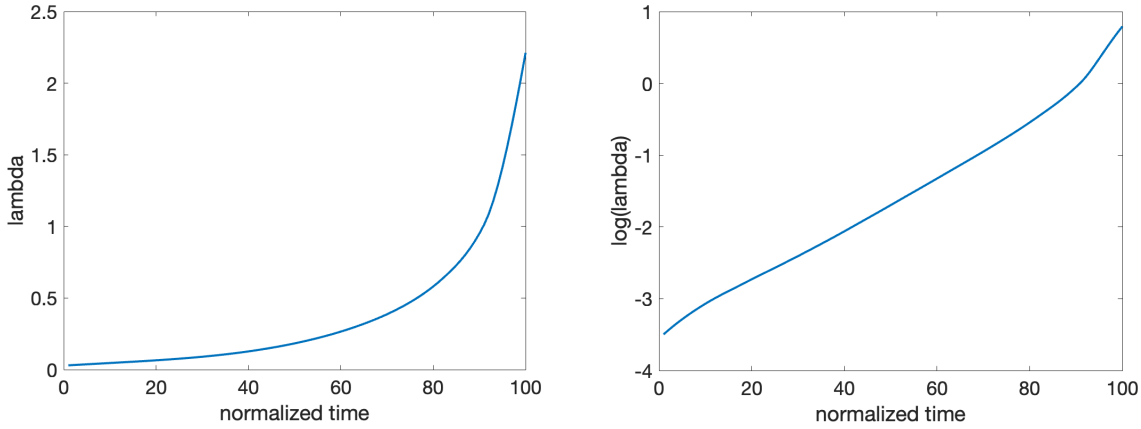


Figure 1: *Trajectory of λ (left) and log(λ) (right) from a Mandarin speaker's production of* [bi].

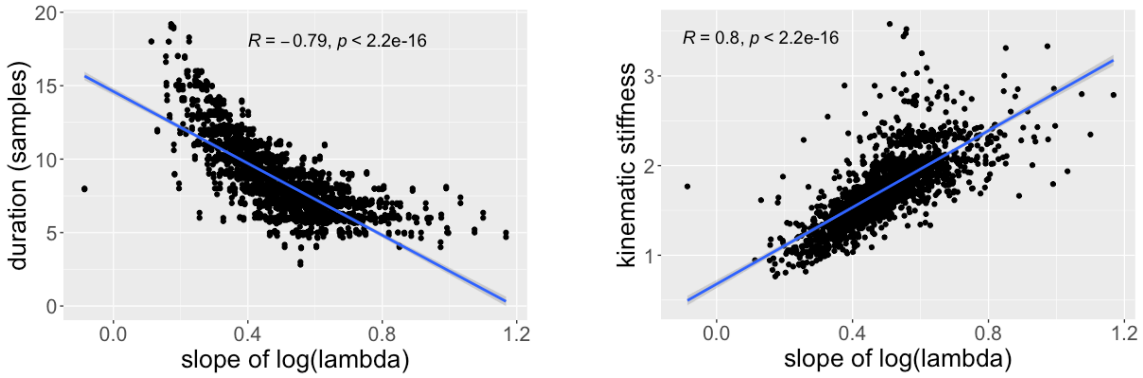


Figure 2: *Relationship between the slope of log(λ) and duration (left) and kinematic stiffness (right).*

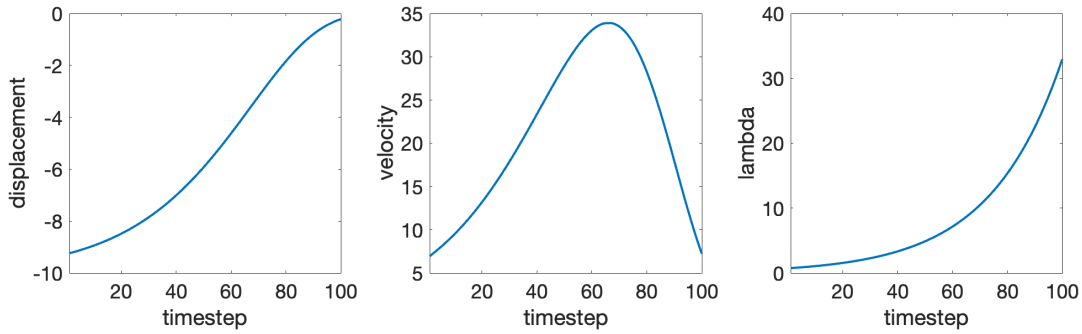


Figure 3: *Displacement (left), velocity (center), and lambda (right) simulated from equations (3) and (4).*

References

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, *26*(2), 173–199. https://doi.org/10.1006/jpho.1998.0071

Kröger, B. J., Schröder, G., & Opgen-Rhein, C. (1995). A gesture-based dynamic model describing articulatory movement data. *The Journal of the Acoustical Society of America*, *98*(4), 1878–1889. https://doi.org/10.1121/1.413374

Ostry, D. J., Cooke, J. D., & Munhall, K. G. (1987). Velocity curves of human arm and speech movements. *Experimental Brain Research*, *68*(1), 37–46. https://doi.org/10.1007/BF00255232

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *The Journal of the Acoustical Society of America*, *77*(2), 640–648. https://doi.org/10.1121/1.391882

Perrier, P., Abry, C., & Keller, E. (1988). Vers une modélisation des mouvements du dos de la langue. *Vers Une Modélisation Des Mouvements Du Dos de La Langue*, *2–1*, 45–63.

Roon, K. D., Hoole, P., Zeroual, C., Du, S., & Gafos, A. I. (2021). Stiffness and articulatory overlap in Moroccan Arabic consonant clusters. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *12*(1), 8. https://doi.org/10.5334/labphon.272

Saltzman, E. L., & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology*, *1*(4), 333–382.

Sorensen, T., & Gafos, A. (2016). The Gesture as an Autonomous Nonlinear Dynamical System. *Ecological Psychology*, *28*(4), 188–215. https://doi.org/10.1080/10407413.2016.1230368

Online compensation of auditory feedback perturbations in speech: an optimal feedback control model of tongue biomechanics

Ny-Tsiky Rakotomalala[1,4], Pierre Baraduc[1,2,4], Pascal Perrier[1,3,4]

[1]*Univ. Grenoble Alpes,* [2]CNRS, [3]*Grenoble INP,* [4]GIPSA-lab

name.surname@grenoble-inp.fr

Introduction. Online responses to perturbations of the auditory feedback have been extensively studied since the pioneering study of Houde and Jordan (1998). A striking finding is that subjects tend to rely preferentially on either the somatosensory or the auditory feedback to adjust speech production (Lametti, Nasir, and Ostry 2012). Another remarkable phenomenon is that even modest online compensation tend to show an asymmetry between up- and down-shifts of the speech formants (Hantzsch, Parrell, and Niziolek 2022). Here we present a modeling analysis of these experiments using GEPPETO-OC This model integrates the neurobiomechanical model of GEPPETO (Payan and Perrier 1997) into a stochastic feedback control framework (Todorov and Jordan 2002), in order to handle high level feedback control during movement execution. This new framework also allows to account for the impact of motor and sensory variability while preserving the idea of optimal effort allocation. We here explore whether the interaction between biomechanical constraints and characteristics of the sensory feedbacks can explain human online audiomotor control.

Methods. Speech goals were considered multimodal: auditory (target region for phonemes defined as ellipsoids in F1/F2 and F3/F2 domains) and proprioceptive (ellipsoids in the 3D space of variables that characterize the tongue shape). For simplicity, the timing of these goal sequence was considered fixed.

GEPPETO-OC does not separate planning and execution: in order to determine the current motor commands, the *optimal controller* computed and adapted the trajectories of the motor commands in real time by constantly minimizing both the neuromuscular effort and the sensory inaccuracy at the phonemic targets as a hybrid cost:

$$C = \sum_{t=t_{\text{current}}}^{T-1} \|[\lambda_(t+1) - \lambda(t)]^+\|^2 + \alpha \sum_{g=g_{\text{current}}}^{N_g} \text{dist}(p(T_g), p_g)$$

T is the total movement duration, λ the motor commands, N_g the number of phonemic targets; p_g the specified sensory values at the phonemic targets ——together with a requirement of final stability (null velocity); $p(T_g)$ contains the actual sensory values at the time T_g set for each target; α is a trade-off parameter.

Motor and sensory signals were assumed corrupted by additive and multiplicative noises, and feedback was assumed delayed by up to 80 ms. An *optimal estimator* (EKF) estimated the state of the system through an internal model of the biomechanics based on an efferent copy of the motor commands, predicted the sensory reafference, and corrected the state estimate as a function of the sensory prediction error.

The original GEPPETO model controlled a finite-element biomechanical model of the tongue. In GEPPETO-OC, to speed up computation, a reduced model was developed. First, the dimensionality of the upper contour of the tongue, described by the position of 16 nodes in the sagittal plane, was reduced to 5D via an autoencoder. Then a LSTM network model of the plant dynamics was trained on thousands of simulations of the finite element model. Last, auditory feedback was computed from the tongue contour using a harmonic model of the vocal tract (Badin and Fant 1984).

Results. To assess whether across-subject differences in sensory precision could explain differential reliance on somatosensory or auditory feedback, we simulated trajectories aimed at the /ɛ/ target from a neutral tongue position in three different conditions: (1) no auditory perturbation; called reference on Figure 1; (2) the auditory feedback of F1 was up-shifted by 125 $mels$ from the onset of the movement; (3) F1 was conversely down-shifted by 125 $mels$.

For these three conditions, we tested two different cases: (a) the auditory signal was more noisy than the somatosensory signal; (b) the somatosensory signal was more noisy than the auditory signal. Each condition was repeated 20 times to compare trial-to-trial variability.

In the (a) case (strong auditory noise), compensation value at the end of the movement is less than 10 $mels$ (8%) in upward and downward shift. In the (b) case, compensation value reach approximately 35 $mels$ (28%). The dynamics of

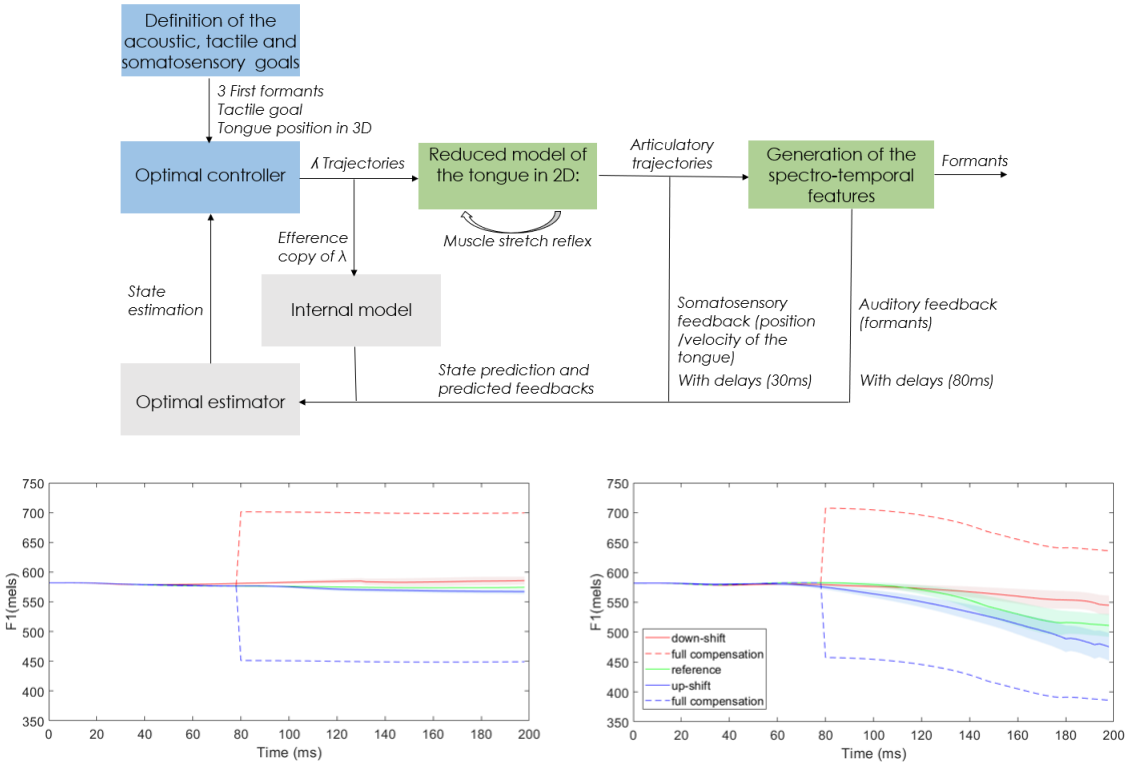


Figure 1: Upper panel : Architecture of GEPPETO-OC. Lower panels: Compensation to auditory perturbation for /ɛ/: left: higher amount of noise in the auditory feedback; right: higher amount of noise in the somatosensory feedbacks. Dashed lines represent a theoretical full compensation level (subtracting the perturbation from the average reference trajectory).

the compensation were also different between (a) and (b), case (a) being close to linear. Compensation of the up-shift and down-shift were also asymmetrical, but differently so in the two cases studied.

Discussion. Our model could reproduce the sensory preference effects by only assuming a difference in sensory precision. Furthermore, the predicted compensation was modest and larger for down-shifts in setting (a), as found experimentally (Hantzsch, Parrell, and Niziolek 2022); this result is likely due to concurrent optimization of precision and effort. In the "auditory preference" case (b), the dynamics of the compensation clearly reflected the asymmetry between an assistive and a resistive perturbation.
Results could also be analyzed in the kinematic domain (tongue contour motion) and in the motor domain (change in muscle activation patterns). This additional data will be presented during the conference.
Future work should compare the effect of a perturbation on either tongue movement (as here) or on stable vowels (static articulatory positions). This would check whether the model is also able to reproduce the experimental observations of a much slower compensation (at around $460\ ms$) under the latter experimental condition.

References.

Badin, P. and G. Fant (1984). "Notes on vocal tract computation". In: *STL QPSR* 2.3, pp. 53–108.

Hantzsch, Lana, Benjamin Parrell, and Caroline A Niziolek (2022). "A single exposure to altered auditory feedback causes observable sensorimotor adaptation in speech". In: *Elife* 11, e73694.

Houde, John F and Michael I Jordan (1998). "Sensorimotor adaptation in speech production". In: *Science* 279.5354, pp. 1213–1216.

Lametti, Daniel R, Sazzad M Nasir, and David J Ostry (2012). "Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback". In: *Journal of Neuroscience* 32.27, pp. 9351–9358.

Payan, Yohan and Pascal Perrier (1997). "Synthesis of VV sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis". In: *Speech communication* 22.2-3, pp. 185–205.

Todorov, Emanuel and Michael I Jordan (2002). "Optimal feedback control as a theory of motor coordination". In: *Nature Neuroscience* 5.11, pp. 1226–1235.

Assessing differences in articulatory-acoustic vowel space in Parkinson's disease phenotypes

Nikki Hoekzema[1], Teja Rebernik[1,2*], Thomas B. Tienkamp[1], Sasha Chaboksavar[1], Valentina Ciot[1], Annetje Gleichman[1], Roel Jonkers[1], Aude Noiray[3], Martijn B. Wieling[1], Defne Abur[1]*

[1]University of Groningen
[2]Vrije Universiteit Brussel
[3]Laboratoire de Psychologie et NeuroCognition, Université Grenoble Alpes
*Both authors contributed equally to this paper
n.hoekzema.1@student.rug.nl, t.rebernik@rug.nl

Introduction. Parkinson Disease (PD) is a neurodegenerative disease presenting with a myriad of symptoms, including muscle rigidity, tremor, slowness of movement, and speech problems. Depending on the dominant symptomatology, individuals with PD (IwPD) can be categorized into either the Tremor Dominant (TD) phenotype, which is primarily characterized by tremor, or Postural Instability Gait Difficulty (PIGD) phenotype, which is characterized by gait disturbance, postural instability, and rigidity (Jankovic, 2008; Stebbins et al., 2013). Relatively little is known about the differences between TD and PIGD phenotypes in articulatory measures of speech. Some prior studies suggest more severe articulatory impairments in PIGD than TD when compared to control speakers on their performance in syllable repetition tasks (Rusz et al., 2023; Tykalová et al., 2020). Another study, based on vowels extracted from a reading passage, suggests a decreased vowel space in IwPD presenting with high bradykinesia and rigidity subscores, but no overall effect of PIGD or tremor subscores (Skrabal et al., 2022). To our knowledge, however, no study has assessed sentence-level (as opposed to word-level) articulatory differences between PIGD and TD phenotypes, and control speakers. Assessing sentence-level articulatory differences allows us to analyze speech across a wider range of vowel productions and is more ecologically valid.

The current study therefore assessed whether there is a difference between PD phenotypes and control speakers in their sentence level vowel production, using the Articulatory Acoustic Vowel Space (AAVS) measure (Whitfield & Goberman, 2014). In addition, we assessed whether other variables, including task (reading vs. elicited speech task), speaker sex, age, cognitive abilities, and hearing status affect AAVS in these three groups. Based on prior studies, we expected IwPD of the PIGD phenotype to show a greater articulatory acoustic vowel impairment (i.e., a reduced AAVS) than control speakers (CS). We additionally expected a larger AAVS in female than male speakers, regardless of group (Whitfield & Goberman, 2014; Houle et al., 2023).

Methods. The study forms part of a larger study, approved by our institutional Medical Ethics Review Board. We report the data of 31 native Dutch IwPD (18 males, 13 females; mean age 69.5 ± 7.7 years) and 29 native Dutch CS (15 males, 14 females; mean age 68.1 ± 7.3 years). All participants completed the Montreal Cognitive Assessment (MoCA) and underwent a hearing screening (without hearing aids). We classified the hearing impairment severity following the Global Burden of Disease Expert Group on Hearing Loss screening (Olusanya et al., 2019), resulting in 23 speakers with none-to-mild hearing impairment (9 CS, 9 TD, 4 PIGD) and 38 speakers with moderate-to-severe hearing impairment (20 CS, 12 TD, 6 PIGD). The tasks were completed while the participants wore their hearing aids and therefore had corrected-to-normal hearing. All IwPD completed Parts I-III of the Movement Disorder Society Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS; Goetz et al., 2008). We classified the motor phenotype according to Stebbins et al. (2013), resulting in 22 TD (11 male, 10 female) and 10 PIGD (7 male, 3 female) IwPD. All IwPD completed the experimental tasks while ON levodopa.

After equipping the participants with a headset microphone (Shure MX153), we asked them to read the North Wind and the Sun Passage ('read' speech) and describe the Cookie Theft picture ('elicited' speech). We subsequently calculated the articulatory acoustic vowel space (AAVS; in mels) following procedures in Whitfield and Goberman (2014). Unlike other vowel space measures (e.g., vowel space area), the AAVS is calculated on the basis of the first and second formants of all voiced segments extracted from running speech, and serves as a measure of overall acoustic vowel spread.

We conducted a linear mixed-effects regression analysis in R version 4.3.1 (R Core Team), using the *lme4* package (Bates et al., 2015). Our hypothesis-testing models included AAVS as the dependent variable, group (TD, PIGD, CS) as the main fixed effect, and sex as an additional fixed effect. We included a by-participant random intercept. In our exploratory analysis, we further assessed the effect of age, task (read vs. elicited speech), hearing impairment (none-to-mild vs. moderate-to-severe impairment) and cognition (MoCA score). We also evaluated whether a two-level group distinction (i.e., PD vs. CS) yielded a better model. Final models were determined via model comparison (using the `anova` function).

The alpha level for rejecting the null hypothesis was set at 0.05. Effect sizes were determined with Cohen's d, which classifies effects as small ($d = 0.2$), medium ($d = 0.5$) or large ($d \geq 0.8$).

Results. Figure 1 visualizes the difference in AAVS between the three groups, separated by sex. In our hypothesis-testing model, there was no significant effect of the PIGD and TD groups on AAVS ($p = 0.1$) compared to CS. There was a significant effect of sex on AAVS ($\beta = 10762$ mel^2, $t = 7.5$, $p < 0.001$, Cohen's $d = 2.0$). The exploratory analysis did not result in a changed model, as including other variables (either separately or in interaction with group) did not yield an improved model (all p's > 0.05). There was therefore no effect of phenotype, age, hearing impairment, cognition or task choice.

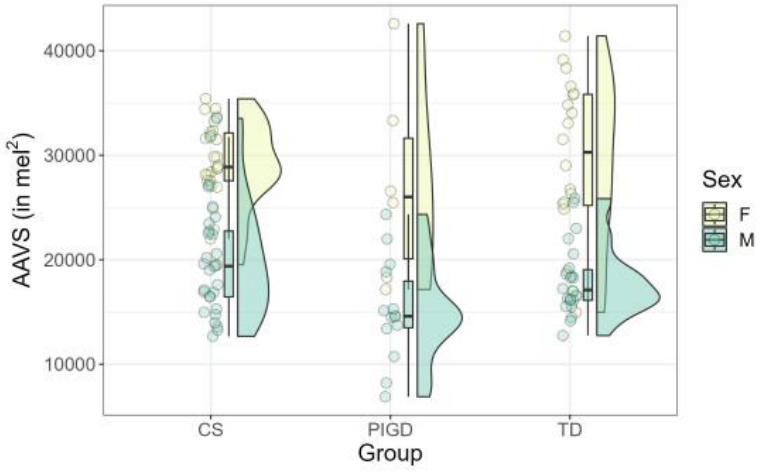


Figure 1: Difference in AAVS (in mel^2) depending on group and sex.

Discussion. Our study results indicate no significant impact of PD phenotype on the AAVS, as there was no difference in AAVS between the control group and IwPD of either the PIGD or TD phenotypes. We likewise did not find any differences between CS and IwPD at the group level. This finding aligns with the results of Houle and colleagues (2023), but conflicts with the results of Whitfield and Goberman (2014), who report a reduced AAVS in IwPD compared to CS. Both prior studies used the Rainbow Passage reading task to assess the AAVS, whereas our study also included an elicited speech task ("Cookie Theft" picture description) next to a read speech task ("The North Wind and the Sun"). The two tasks were comparable in terms of the AAVS, indicating that choosing a more ecologically valid ('elicited speech') task is a suitable choice for researchers wishing to evaluate differences in the articulatory-acoustic vowel space. Following previous studies (Whitfield & Goberman, 2014; Houle et al., 2023), female speakers exhibited a significantly larger AAVS than male speakers. While our study includes a limited participant sample (10 PIGD compared to 22 TD and 29 CS participants), the results provide a first glimpse into sentence-level articulation of speakers of different IwPD phenotypes.

References

Bates D., Mächler M., Bolker B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., …, LeWitt, P. A., Nyenhuis, D., Olanow, C. W., Movement Disorder Society UPDRS Revision Task Force (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society, 23*(15).

Houle, N., Feaster, T., Mira, A., Meeks, K., & Stepp, C. E. (2023). Sex Differences in the Speech of Persons with and without Parkinson's Disease. *American journal of speech-language pathology*, 1–21.

Jankovic, J. (2008). Parkinson's disease: Clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry,* 79(4), 368–376.

Olusanya, B. O., Davis, A. C., & Hoffman, H. J. (2019). Hearing loss grades and the International classification of functioning, disability and health. *Bulletin of the World Health Organization*, 97(10), 725–728.

R Core Team (2023). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rusz, J., Krupička, R., Vítečková, S., Tykalová, T., Novotný, M., Novák, J., Dušek, P., & Růžička, E. (2023). Speech and gait abnormalities in motor subtypes of de-novo Parkinson's disease. *CNS Neuroscience & Therapeutics, 29*(8), 2101–2110.

Stebbins, G. T., Goetz, C. G., Burn, D. J., Jankovic, J., Khoo, T. K., & Tilley, B. C. (2013). How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: comparison with the unified Parkinson's disease rating scale. *Movement disorders: official journal of the Movement Disorder Society*, 28(5), 668–670.

Skrabal, D., Rusz, J.., Novotny, M., Sonka, K., Ruzicka, E., Dusek, P., & Tykalova, T. (2022). Articulatory undershoot of vowels in isolated REM sleep behavior disorder and early Parkinson's disease. *NPJ Parkinson's disease.* https://doi.org/10.1038/s41531-022-00407-7

Tykalová, T., Rusz, J., Švihlík, J., Bancone, S., Spezia, A., & Pellecchia, M.T. (2020). Speech disorder and vocal tremor in postural instability/gait difficulty and tremor dominant subtypes of Parkinson's disease. *Journal of Neural Transmission, 127*(9), 1295–1304.

Whitfield, J. A., & Goberman, A. M. (2014). Articulatory-acoustic vowel space: application to clear speech in individuals with Parkinson's disease. *Journal of communication disorders*, 51, 19–28.

Effects of expectedness and clarity of speech auditory feedback on perception and motor control

Abigail R. Bradshaw[1], Clément Gaultier[1], Clare Press [2,3], Matt Davis[1]

[1]MRC Cognition and Brain Sciences Unit, University of Cambridge
[2] Experimental Psychology, University College London
[3]Wellcome Centre for Human Neuroimaging, University College London
abbie.bradshaw@mrc-cbu.cam.ac.uk

Introduction. Prediction has been proposed to play a role in both perception and action; however, there has been some debate over the specific mechanism by which predictions are integrated with sensory input, and whether this is consistent across both perception in passive settings and perception of self-generated movement effects. Traditionally, while a sharpening mechanism has been dominant in the perception literature (in which representation of expected sensory input is enhanced), a prediction error cancellation mechanism has dominated the field of motor control (in which representation of expected sensory input is suppressed). One method to interrogate the operation of such predictive mechanisms at the behavioural level is to test whether prior expectations have an enhancing or suppressing effect on perceptual outcomes. Previous evidence has found that perception of the clarity of speech produced by others was enhanced by accurate prior expectations (Sohoglu et al., 2014); interestingly however, multivariate fMRI and MEG evidence with the same paradigm has reported a unique neural signature of prediction error coding in which the effect of expectedness on decoding success (i.e. representation strength) was dependent on the level of perceptual clarity (Blank et al., 2018; Blank & Davis, 2016; Sohoglu & Davis, 2020). The current study aimed to contribute behavioural evidence to this debate by investigating whether perception of self-generated speech during speech production is enhanced or suppressed by accurate expectations.

Methods. To test this, we manipulated (1) the expectedness and (2) the clarity of real-time speech auditory feedback during the production of single words. Expectedness was manipulated using real-time perturbation of speech formants (resonant frequencies that determine vowel sounds); specifically, the first formant was shifted up or down on randomly occurring trials (as in a typical compensation paradigm). Clarity was manipulated using real-time noise vocoding, a technique that allows for parametric degradation of the level of spectral detail present in a speech signal (Shannon et al., 1995). These manipulations resulted in a 2x3 design, with two levels of expectedness (altered or unaltered speech feedback) and three levels of clarity (high, medium and low), which occurred randomly across trials. We measured the effect of these two manipulations on three outcome measures: two explicit measures of perceptual experience (rated clarity of speech feedback and detection of feedback perturbations) and one implicit measure of motor control (trial-by-trial compensation for the random perturbation). We predicted that ratings of clarity would be higher for unaltered versus altered speech, reflecting an enhancing effect of prior expectations on perception, in line with findings from passive speech perception. This effect of expectedness may decrease with decreasing clarity (due to poorer discrimination between altered and unaltered utterances) or may increase with decreasing clarity (due to an increase in the relative influence of the prior expectation with noisier sensory feedback, as in Bayesian accounts). We predicted that there would be a significant effect of clarity on compensation, but no significant effect of perturbation detection (i.e. whether participants reported hearing the perturbation or not).

Results. Pilot data ($n = 3$) results suggest that perceptual outcomes for the self-voice during active speech production are enhanced when speech feedback matches expectations; that is, participants reported higher perceived clarity for productions with unaltered speech feedback than productions with altered speech feedback. This effect of expectedness further appears to decrease with decreasing clarity. The pilot data results also suggest that expectedness may have different effects on rated clarity according to whether participants report detecting the perturbation or not (when controlling for the clarity condition); specifically, higher clarity was reported for the unaltered than the altered condition when participants reported no perturbation, whereas the reverse was true when participants reported hearing a perturbation. Further planned analyses with the full dataset will test for the effect of both clarity and perturbation detection on implicit compensation.

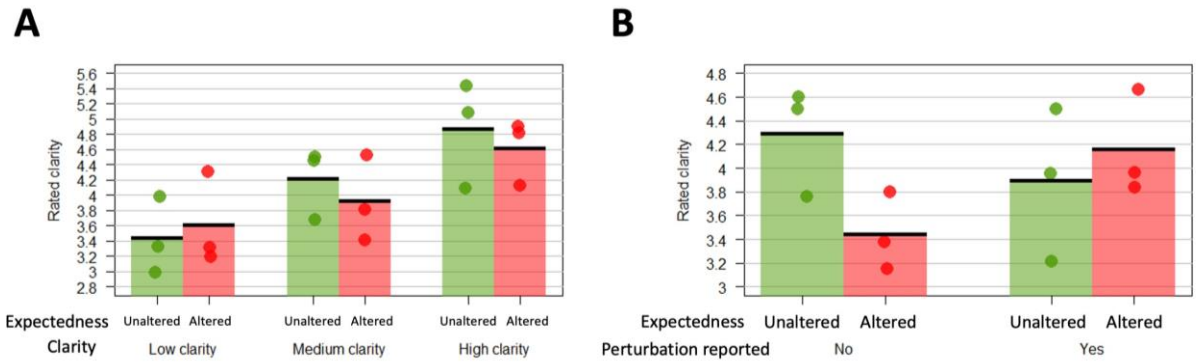


Figure 1: *(A) Effect of expectedness (unaltered or altered speech auditory feedback) and clarity on rated clarity. (B) Effect of expectedness on rated clarity, according to whether perturbation was reported or not (shown for the medium clarity condition).*

Discussion. Overall, the results from our preliminary pilot data suggest that perception of the self voice during speech production shows an enhancing effect of prior expectations. After confirmation of this in a larger sample ($n = 20$), we plan to transfer this paradigm into fMRI to test whether the neural mechanism underlying this perceptual sharpening effect relies on an intermediate stage in which prediction errors are calculated in posterior auditory cortex, using multivariate decoding methods. This will test for the presence of a 'neural signature' of prediction error coding demonstrated previously during passive speech perception (Sohoglu & Davis, 2020); specifically, an interaction was found in which decoding success increased with increasing clarity when speech was unexpected, but decreased with increasing clarity when speech was expected.

References

Blank, H., & Davis, M. H. (2016). Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. *PLOS Biology*, *14*(11), e1002577.

Blank, H., Spangenberg, M., & Davis, M. H. (2018). Neural Prediction Errors Distinguish Perception and Misperception of Speech. *Journal of Neuroscience*, *38*(27), 6076–6089. https://doi.org/10.1523/JNEUROSCI.3258-17.2018

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, *270*(5234), 303–304. https://doi.org/10.1126/science.270.5234.303

Sohoglu, E., & Davis, M. H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. *ELife*, *9*. https://doi.org/10.7554/eLife.58077

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2014). Top-Down Influences of Written Text on Perceived Clarity of Degraded Speech. *Journal of Experimental Psychology- Human Perception and Performance*, *40*(1), 186–199. https://doi.org/10.1037/a0033206

Quantifying the Lombard Effect in Noisy Environments

Jackie S. Kim[1], Alan Bush[2,3], R. Mark Richardson[2,3], Matteo Vissani[2,3], Frank Guenther[1,4,5]

[1] *Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA*
[2] *Department of Neurosurgery, Massachusetts General Hospital, Boston, MA*
[3] *Harvard Medical School, Boston, MA*
[4] *Department of Biomedical Engineering, Boston University, Boston, MA*
[5] *Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA*

Author 1: jskim1@bu.edu, Author 2: alan.bush@mgh.harvard.edu

Introduction. The Lombard effect is a reflex phenomenon where speakers modulate their articulatory movements and acoustic speech properties to increase speech intelligibility (i.e., improve the perceived signal-to-noise ratio) to listeners in noisy environments. In speech therapy, the Lombard reflex may be a useful technique to elicit higher vocal intensity in people with hypophonia (Stathopoulos et al., 2014). Approximately 90% of people with Parkinson's disease (**PD**), a disorder of the basal ganglia, present with hypophonia and hypokinetic dysarthria (Liotti et al., 2002; Weiss et al., 2023). The basal ganglia is strongly implicated in the motor aspects of speech production, with evidence of subthalamic nucleus (**STN**) involvement in cognitive processes for speech such as voluntary gain modulation, motor learning, and action selection/suppression (Guenther & Hickock, 2015; Chrabaszcz et al., 2019). Despite the significant impact of PD on speech communication, the role of the basal ganglia in involuntary speech gain modulation is poorly understood. Thus, we leveraged the unique opportunity provided by deep brain stimulation (**DBS**) implantation surgeries in PD patients to study STN and globus pallidus internus (**GPi**) activity during speech. Specifically, the Lombard effect was utilized to induce speech gain modulations in people with PD. Most studies on the Lombard effect present noise to participants through headphones. Although this experimental approach allows speech recordings with minimal noise, the occlusion effect created by blocking the ear canals can disrupt the speakers' self-monitoring auditory feedback through bone conduction and does not accurately reflect real-world speaking environments (Vaziri et al., 2019). For this study, participants produced sentences in either a quiet or noisy condition with multi-speaker babble presented through a loudspeaker, resulting in a more realistic perception of their speech without any aural occlusion. Headphones were not used in this setting due to the complex setup of neurosurgery in the operating room and ethical reasons to maximize communication between patients and the medical team. However, this method introduced new challenges in extracting the recorded speech signal from the background noise of a noisy operating room (Srivastava et al., 2021). Therefore, the behavioral results in the experiment to accurately quantify the Lombard effect will be presented.

Methods. Participants included 12 native English speakers with PD (10M/2F; mean age: 67.1 $\pm$ 9.8 years) undergoing DBS implantation for PD. The patients underwent awake stereotactic neurosurgery for implantation of DBS electrodes in the STN (n=7) or GPi (n=5). Dopaminergic medication was withdrawn the night prior to surgery. All patients provided informed consent to participate in the study. All procedures were approved by Massachusetts General Hospital Institutional Review Board and performed at the Massachusetts General Hospital.

For the task, participants repeated 10 different Harvard sentences with or without noise (multi-speaker babble) present in the background. Stimuli were presented auditorily through a speaker in either a quiet or noisy condition and orthographically displayed on a screen. Prior to surgery, the speaker audio was calibrated from the position of the participant's ears at a sound pressure level (**SPL**) of 70 decibels (dB) to achieve a Lombard effect of adequate magnitude. The speech signal was recorded with a directional microphone targeted at 15 centimeters from the participant's mouth. Each run of the task took approximately eight minutes. The acoustic echo cancellation (**AEC**) algorithm implemented in the speexdsp library was applied to audio recordings to remove stimulus audio captured by the microphone (Valin, 2016). The sound files were then processed through the *Penn Phonetics Lab Forced Aligner* toolkit, which automatically aligns the phonemes in each sentence with the corresponding section of the spectrogram (Yuan & Lieberman, 2008). A certified speech-language pathologist reviewed each output file to check for any misalignments and corrected them if necessary. After the sound files were processed, acoustic features including fundamental frequency (F0), formants (F1 and F2), intensity, and vowel centralization were extracted for further analysis using Praat and Matlab. The mean intensity of five beeps played at 1 kilohertz used to calibrate the microphone before the task was used as the absolute SPL of the speech signal, then applied to the quiet condition of the Lombard task. The values are used to report the produced speech intensity in absolute dB SPL using Praat. The root mean square (**RMS**) of the intertrial gaps and vowels before and after AEC was calculated in pascals. Speech intensity was calculated as $SPL(dB) = 20 \times \log_{10}(RMS)$, the measure of the pressure fluctuation in a sound wave relative to a reference pressure (Laukli & Burkard, 2015). The SPL of the participants' speech was compared in both conditions (i.e., quiet and noisy) before and after AEC.

Results. The Lombard reflex was evident in the PD patients during the awake DBS implantation surgery, with no significant group differences based on the site of electrode implantation. Incorporating the mean intensity of the beeps recorded prior to the experiment was critical to calibrate the intensity of the speech signal. The mean SPL was 67.6 ± 4.3dB in the quiet condition and 70.9 ± 4.0dB in the noisy condition. The mean increase in vocal intensity from the quiet to noisy condition was 3.6dB. The mean SPL for the intertrial gap in the quiet condition was 51.7dB prior to applying the AEC algorithm and 49.2dB afterward, with a 2.6dB difference. In the noisy condition, the mean SPL for the intertrial gap was 64.5dB in prior to applying the algorithm and 52.4 afterward, with a significant 12.3dB difference of denoising the signal. The mean intensity of the beeps was 73.1dB. The acoustic methods utilized in the study precisely quantify the Lombard effect by employing an acoustic echo cancellation algorithm on the speech task audio, applying absolute volume calibration, and comparing the intensities between the intertrial gaps and speech signals in both conditions. Systematic acoustic methods should be employed when analyzing the Lombard effect in noisy environments to prevent inadvertently capturing noise within the speech signal.

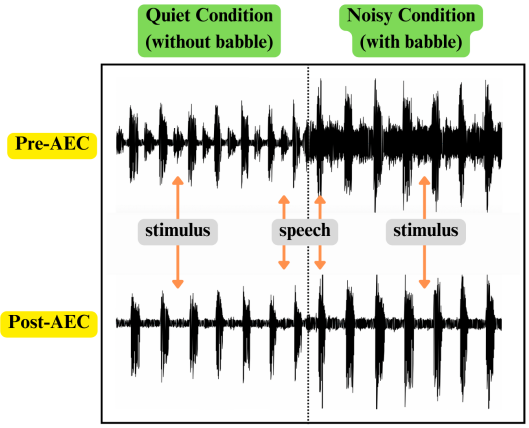


Figure 1: *Quiet vs. noisy condition during Lombard task before and after acoustic echo cancellation (AEC)*

Discussion. The current study shows that the Lombard reflex can be robustly induced and quantified in an intraoperative setting with natural auditory feedback. The absence of headphones reduces the occlusion effect and makes the results more comparable and applicable to everyday acoustic environments. Although there are numerous studies on the Lombard effect, the underlying mechanisms of how speakers with PD modulate their speech is still unclear. These results will be combined with simultaneous intracranial recordings to further identify key neural correlates of speech gain modulation. Broadening our understanding of how subcortical regions of the brain control speech can contribute to developing more functional speech therapy goals for people with motor speech impairments from different movement disorders.

References

Chrabaszcz, A., Neumann, W. J., Stretcu, O., Lipski, W. J., Bush, A., Dastolfo-Hromack, C. A., Wang, D., Crammond, D. J., Shaiman, S., Dickey, M. W., Holt, L. L., Turner, R. S., Fiez, J. A., & Richardson, R. M. (2019). Subthalamic Nucleus and Sensorimotor Cortex Activity During Speech Production. *The Journal of neuroscience: The Official Journal of the Society for Neuroscience*, *39*(14), 2698–2708. https://doi.org/10.1523/JNEUROSCI.2842-18.2019

Guenther, F. H., & Hickok, G. (2015). Neural models of motor speech control. In Neurobiology of language (pp. 725–740). Academic Press. https://doi.org/10.1016/B978-0-12-407794-2 .00058-4

Kleiner-Fisman, G., Herzog, J., Fisman, D. N., Tamma, F., Lyons, K. E., Pahwa, R., Lang, A. E., & Deuschl, G. (2006). Subthalamic nucleus deep brain stimulation: Summary and meta-analysis of outcomes. Movement Disorders, 21(S14), S290–S304. https:// doi.org/10.1002/mds.20962, PubMed: 16892449

Laukli, E., & Burkard, R. (2015). Calibration/Standardization of Short-Duration Stimuli. *Seminars in hearing*, *36*(1), 3–10. https://doi.org/10.1055/s-0034-1396923

Liotti, M., Ramig, L. O., Vogel, D., New, P., Cook, C. I., & Fox, P. T. (2002). Hypophonia in Parkinson disease: Neural correlates of voice treatment with LSVT revealed by PET. In 7th International Conference on Spoken Language Processing, ICSLP 2002 (pp. 2477–2480). International Speech Communication Associa- tion. https://doi.org/10.21437/ICSLP.2002-645

Srivastava, P., Shetty, P., Shetty, S., Upadya, M., & Nandan, A. (2021). Impact of Noise in Operating Theater: A Surgeon's and Anesthesiologist's Perspective. *Journal of pharmacy & bioallied sciences*, *13*(Suppl 1), S711–S715. https://doi.org/10.4103/jpbs.JPBS_656_20

Stathopoulos, E. T., Huber, J. E., Richardson, K., Kamphaus, J., DeCicco, D., Darling, M., Fulcher, K., & Sussman, J. E. (2014). Increased vocal intensity due to the Lombard effect in speakers with Parkinson's disease: simultaneous laryngeal and respiratory strategies. *Journal of communication disorders*, *48*, 1–17. https://doi.org/10.1016/j.jcomdis.2013.12.001

Valin, J. M. (2016). Speex: A free codec for free speech. *arXiv preprint arXiv:1602.08668*.

Vaziri, Ghazaleh & Giguère, Christian & Dajani, Hilmi. (2019). Evaluating noise suppression methods for recovering the Lombard speech from vocal output in an external noise field. International Journal of Speech Technology. 22. 10.1007/s10772-018-09564-8.

Yuan, J. And Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *In Proceedings of Acoustics 2008,* 5687-5690.

Weiss, A. R., Korzeniewska, A., Chrabaszcz, A., Bush, A., Fiez, J. A., Crone, N. E., & Richardson, R. M. (2023). Lexicality-modulated influence of auditory cortex on subthalamic nucleus during motor planning for speech. Neurobiology of Language, 4(1), 53–80. https://doi.org/10.1162 /nol_a_00086

The Effect of Speaking Style on the Articulatory-Acoustic Vowel Space in Individuals with Tongue Cancer Before and After Surgical Treatment

Thomas B. Tienkamp[1], Teja Rebernik[1], Raoul Buurke[1], Katharina Polsterer[1], Rob J.J.H. van Son[2,3], Martijn Wieling[1], Max J.H. Witjes[1], Sebastiaan A.H.J. de Visscher [1], Defne Abur[1]

[1]*University of Groningen,* [2]*Netherlands Cancer Institute,* [3]*University of Amsterdam*

t.b.tienkamp@rug.nl

Introduction. Surgical intervention for tongue cancer often has a negative impact on tongue mobility (Lazarus et al. 2014). This reduced mobility may lead to more centralised speech where the acoustic distance between phonemes becomes smaller, which may consequently affect speech intelligibility. Studies that assess the effect of surgical treatment on speech acoustics often use isolated words or sentences. These stimuli typically elicit a best-effort attempt and may not be reflective of a speaker's typical daily communication because read speech results in a larger vowel space than more spontaneously elicited speech (Nakamura, Iwano, and Furui 2008). A better understanding of how speaking style affects vowel acoustics before and after treatment may aid in the development of a standardised speech assessment for individuals with tongue cancer. The objective of this study was therefore to assess the effect of speaking style on the comprehensive acoustic vowel space in individuals undergoing surgical treatment for oral cancer. This was done by measuring the Articulatory-Acoustic Vowel Space (AAVS) across a reading passage and across more spontaneously elicited speech in individuals before and after surgical treatment for tongue cancer. An advantage of the AAVS over purely vowel-based metrics (e.g., the Vowel Space Area, or VSA) is that the AAVS may be computed over running speech and takes all vowels into account, thus preserving some form of ecological validity.

Methods. The present study is part of a larger project approved by the institution's Medical Ethics Review Board. All participants provided written informed consent prior to their participation. Seven native speakers of Dutch (four male, three female) with a mean age of 62 years (SD = 13 years, range = 42 to 77 years) participated in this study a few days before surgery, and approximately six months after surgery for T1 (n = 4), T2 (n = 2), or T3 (n = 1) tongue carcinomas (six lateral, one midsagittal). Two participants received a flap reconstruction. Other participants were locally closed.
Participants read 15 ten-word sentences that included the full phoneme distribution of Dutch (Luts et al. 2014). Participants were fitted with an omni-directional microphone (Shure MX-153) with a seven cm mouth-to-microphone distance. To elicit more spontaneous speech, participants were also asked to describe two pictures in detail: The Cookie Theft picture, and The Cat Rescue picture (Goodglass, Kaplan, and Weintraub 2001; Nicholas and Brookshire 1993). The AAVS was calculated for each speaker at each time point, and for each speaking style. The AAVS was measured on a mel scale using continuous F_1 and F_2 formant tracks of all voiced segments, based on Whitfield and Goberman (2014).
All recordings and formant tracks were checked prior to data analysis. The data was statistically analysed using linear mixed-effects regression in R 4.2.0 (R Core Team 2023; Bates et al. 2015). Our hypothesis model included the AAVS as a function of time (pre vs. post surgery) in interaction with style (read speech vs. elicited speech), together with a by-speaker random intercept. We further assessed the influence of speaker sex and articulation rate (syllables / speaking time) in an exploratory manner. The articulation rate was calculated using a Praat script by De Jong and Wempe (2009). All numerical variables were centered around the mean.

Results. An overview of the AAVS values per speaking style and time point is provided in Figure 1. The AAVS at the 6-month follow-up was not significantly smaller compared to baseline (β = -124 mel^2, $p = .91$). Elicited speech yielded a significantly smaller AAVS when compared to read speech (β = -3208 mel^2, $p < .01$). There was no significant interaction between time and style (β = 666 mel^2, $p = .20$). A fixed effect of sex indicated that males had a smaller AAVS compared to females (β = -12857 mel^2, $p < .01$). Lastly, a fixed effect of articulation rate indicated that those with a faster articulation rate had a larger AAVS (β = 6039 mel^2, $p < .05$).

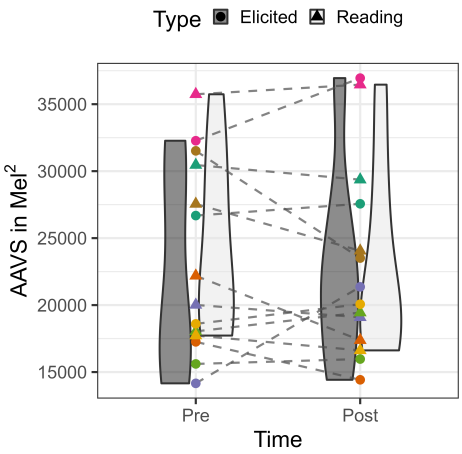


Figure 1: Articulatory-acoustic vowel space (AAVS) per time point and speaking style. Different colours represent individual speakers, different shapes represent the speaking styles (circles = elicited, triangles = reading).

Discussion. We assessed the effect of speaking style on the articulatory-acoustic vowel space (AAVS) of individuals before and after surgical treatment for tongue cancer. The results suggest that the surgical intervention did not lead to an overall vowel space reduction for the speakers included in this study. When comparing our findings to the results from typical Dutch speakers in Hoekzema et al. (Submitted), it seems that the participants in our study with oral cancer had typical articulatory-acoustic vowel spaces before surgical intervention. On average, we found that speakers with higher articulation rates have a larger AAVS. The results of our study further suggest that the AAVS is able to capture subtle differences induced by speaking style. Whereas previous work showed that the AAVS was related to clear speech (larger AAVS values were found for clear speech compared to typical speech (Whitfield and Goberman 2014)), we extend these findings by showing that elicited speech resulted in a smaller AAVS as compared to read speech. The effect of speaking style on AAVS did not change as a result of surgery for the speakers in our study. If the goal is to quantify the effect of surgical treatment for tongue cancer on the vowel space, our results suggest that a more ecologically valid task (i.e., picture description) would be suitable provided that the difference between read and elicited speech did not change over time. We do note, however, that this is based on a small group level assessment and individual patterns may exist.

References.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

De Jong, Nivja H. and Ton Wempe (2009). "Praat script to detect syllable nuclei and measure speech rate automatically". In: *Behavior Research Methods* 41.2, pp. 385–390. DOI: 10.3758/BRM.41.2.385.

Goodglass, Harold, Edith Kaplan, and Sandra Weintraub (2001). *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA.

Hoekzema, Nikki, Teja Rebernik, Thomas B. Tienkamp, Sasha Chaboksavar, Valentina Ciot, Annetje Gleichman, Roel Jonkers, Aude Noiray, Martijn B. Wieling, and Defne Abur (Submitted). "Assessing differences in articulatory-acoustic vowel space in Parkinson's disease phenotypes". In: *13th International Seminar on Speech Production*.

Lazarus, C. L., H. Husaini, A. S. Jacobson, J. K. Mojica, D. Buchbinder, D. Okay, and M. L. Urken (2014). "Development of a New Lingual Range-of-Motion Assessment Scale: Normative Data in Surgically Treated Oral Cancer Patients". In: *Dysphagia* 29.4, pp. 489–499. DOI: 10.1007/s00455-014-9534-9.

Luts, Heleen, Sofie Jansen, Wouter Dreschler, and Jan Wouters (2014). "Development and normative data for the Flemish/Dutch Matrix test". In: *Unpublished Article*.

Nakamura, Masanobu, Koji Iwano, and Sadaoki Furui (2008). "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance". In: *Computer Speech & Language* 22.2, pp. 171–184. DOI: 10.1016/j.csl.2007.07.003.

Nicholas, Linda E and Robert H Brookshire (1993). "A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia". In: *Journal of Speech, Language, and Hearing Research* 36.2, pp. 338–350.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Whitfield, Jason A. and Alexander M. Goberman (2014). "Articulatory–acoustic vowel space: Application to clear speech in individuals with Parkinson's disease". In: *Journal of Communication Disorders* 51, pp. 19–28. DOI: 10.1016/j.jcomdis.2014.06.005.

Aryepiglottic trilling in Mehweb: acoustics and variability

Ekaterina Shepel[1], Alexandre Arkhipov[2], Michael Daniel[3], Alexander Shiryaev[4]

[1]HSE University
[2]Universität Hamburg, Germany
[3]Collegium de Lyon / Laboratoire Dynamique du Langage, France
[4]Independent researcher, Singapore

katerina11780@gmail.com, alexandre.arkhipov@uni-hamburg.de,
misha.daniel@gmail.com, ashiryaev87@gmail.com

Introduction. We investigate the acoustic properties and their intra- and inter-speaker variation for aryepiglottic trilling in epilaryngeal fricatives [ħ] and [ʜ] in Mehweb (Glottocode: mege1234; Dargwa branch of the East Caucasian, or Nakh-Daghestanian, family). Below, we use 'epilaryngeal' as a term covering both 'epiglottal' and 'pharyngeal'.
In Mehweb, the contrasting post-uvular segments are represented by glottal /ʔ/ and /h/ and epilaryngeal /ħ/. A salient feature in Mehweb is pharyngealization. Its presence is a lexical property of morphemes; it surfaces on both vowels and consonants but is mostly confined to syllables with either epilaryngeals or uvulars. Segments [ʔˤ] and [ʜ] most often occur under pharyngealization and - at least in this case - are claimed by Moroz (2019) to be allophones of /ʔ/ and /ħ/ respectively. Here, we focus on epilaryngeal fricatives, whose main acoustic properties we discuss in Arkhipov et al (2023). AE trilling was previously mentioned as a potential additional factor of differentiating between the plain [ħ] and pharyngealized epilaryngeal [ʜ] but has not been discussed in detail.
The nature of AE trilling is described in Esling (2005), suggesting that AE trilling becomes available in laryngeally constricted contexts: "[o]nce compression of the constrictor mechanism is tight enough, the aryepiglottic folds can trill against the epiglottal surface". Therefore epiglottal [ʜ] was qualified as an enhanced version of the pharyngeal [ħ], whereby AE trilling is one of the available phonetic options of enhancement. Based on a pilot data sample, Arkhipov et al (2019) have shown that in Mehweb, AE trilling "may appear, depending on the speaker, in non-pharyngealized roots only…, or sporadically in both contexts…, or not at all.", thus suggesting, pace Esling (2005), that laryngeally less constricted (non-pharyngealized) context can also be more favorable to AE trilling. In this paper, we provide more robust evidence for this claim.

Data. The recordings were made in the village of Mehweb in 2019 and 2022 by the third author. In sum, data of 11 speakers (4 women and 7 men) were collected, some of whom were recorded both in 2019 and 2022. Most speakers were born between 1955 and 1971, one male speaker was born in 1994.
The original data included stimuli with epilaryngeal stop [ʔˤ] and fricatives [ħ], [ʜ], glottal stop [ʔ] and fricative [h]. All sounds were recorded in word-initial, medial, and final position across several vocalic contexts. In this study we focus on epilaryngeal fricatives alone. The first list used in 2019 included 35 stimuli. A longer list used in 2022 includes 134 stimuli, though the exact set of the recorded stimuli varies slightly between speakers. For each speaker, the order of the stimuli was randomized. Russian translation of words was given by the interviewer, sometimes followed by a native prompt where the word was difficult to recollect. In 2019, each word was produced by speakers three times in isolation; in 2022 – four times in isolation and once in a carrier phrase freely composed by the consultant. In this study we do not control for the difference between isolated and phrasal realizations. Audio was recorded with a Sennheiser HSP4-EW headset and Olympus DM-901 voice recorder as Linear PCM at 48 kHz/16 bit.

Methods. The recordings were manually segmented into phones and labeled in Praat. In carrier phrases, only words containing target sounds were annotated. For the target sounds, in case of presence of AE trilling, each pulse was marked on a point tier. The presence of trilling was first estimated auditorily and by visually inspecting the spectrogram. To assist in detection and markup of trilling pulses, we used intensity curves calculated from band-pass filtered sound. Depending on the speaker, various bands were tried, most often 8 to 14 kHz or 10 to 14 kHz. Measures extracted from the annotation included number of pulses, normalized time of the first and last pulses and relative duration of the trilling compared to that of the sound, median period (and frequency) of the pulses, and mean intensity of pulses (difference between peaks and valleys).

Results. AE trilling is marked by periodic increase of energy throughout the spectrum of the fricative at ca. 40–60 Hz frequency (in two female speakers, the frequency is higher and reaches 100 Hz on some tokens). The increase is usually best visible in spectrograms in the region between 8 kHz and 14 kHz (Fig. 1). Based on our analysis, we can divide all speakers into three groups. In the first group (one male and one female speaker; interestingly, they are husband and wife), AE trilling occurs most frequently. In the male speaker (Ab), 42 out of the total 328 epilaryngeal tokens contained AE trilling; in the female speaker (Zm), 139 out of 341 tokens. Notably, all but three AE trilling tokens in the male speaker

were found in 2019 data. In the recordings from the second group (six speakers), AE trilling was found in less than 20 realizations of the target sounds per speaker. In the last group (three speakers), AE trilling was not attested at all.
In the two speakers with the high occurrence of AE trilling, it is found ca. 10 times more often on the plain [ħ] than on the pharyngealized [ʜ]. In the data from the second group, AE trilling is also more frequent on [ħ] than on [ʜ] (44 and 16 occurrences for all speakers, respectively).
Across all data, AE trilling has on average less intensity on the pharyngealized [ʜ] than on the plain [ħ].
In word-medial position, AE trilling tends to continue longer relative to the duration of the fricative. Word-initially, less than 50% towards the end of the sound is usually trilled. As for the specific vowel preferences, there is a tendency for AE trilling to occur before [u] and [i] vowels. However, this is not a rule, as few tokens were found where AE trilling does not appear in fricatives before [u].
Fricatives with AE trilling are never voiced in our data, while voicing without AE trilling does occur.

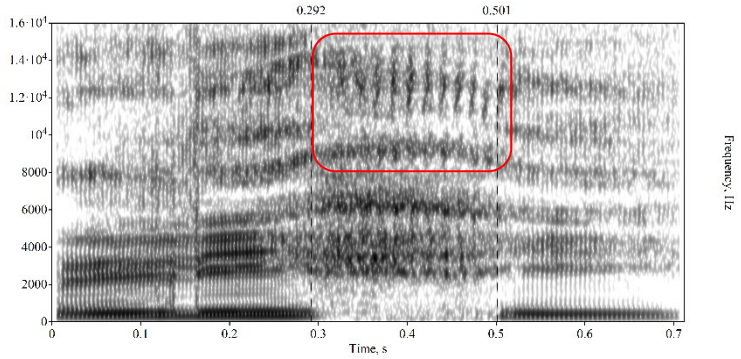


Figure 1. AE trilling in *liħi* ‘ear’, Speaker Zm.

Table 1: *AE trilling on plain [ħ] and pharyngealized [ʜ] in recordings from 2019 and 2022 (speakers Zm and Ab): number of epilaryngeal tokens with trilling / total tokens, median frequency and relative duration of trilling*

Spk / Year	Sound	Word position: Initial			Medial			Final			Total
		Tokens	Freq	RelDur	Tokens	Freq	RelDur	Tokens	Freq	RelDur	Tokens
Zm 2019	plain	34 / 50	60	0,40	11 / 14	63	0,56	0 / 17	–	–	45 / 81
	phar	4 / 11	56	0,34	3 / 15	58	0,18	0 / 10	–	–	7 / 36
Zm 2022	plain	54 / 65	57	0,48	27 / 42	58	0,63	2 / 22	52	0,39	83 / 129
	phar	3 / 22	52	0,26	1 / 45	43	0,47	0 / 28	–	–	4 / 95
Ab 2019	plain	13 / 52	54	0,43	10 / 14	48	0,60	16 / 24	40	0,57	39 / 90
	phar	0 / 9	–	–	0 / 13	–	–	0 / 9	–	–	0 / 31
Ab 2022	plain	2 / 62	49	0,35	0 / 35	–	–	1 / 20	67	0,12	3 / 117
	phar	0 / 26	–	–	0 / 40	–	–	0 / 24	–	–	0 / 90

Discussion. The plain epilaryngeal fricative [ħ] and the pharyngealized epilaryngeal fricative [ʜ] behave differently in terms of AE trilling, with the presence and extent of AE trilling varying strongly between speakers and also across tokens produced by the same speaker. Only in the recordings from 2 out of 11 speakers is AE trilling relatively frequent. In these cases, it occurs mostly on the non-pharyngealized [ħ]. In the other 6 speakers who produced AE trilling it is present more often on [ħ] than on [ʜ]. Presumably, while AE trilling might need some degree of laryngeal constriction to occur (Esling 2005, Esling et al. 2019), excessive constriction associated with pharyngealization can impede the vibrations, and is likely to inhibit AE trilling in pharyngealized contexts in Mehweb.
In our data, AE trilling appears mostly on fricatives followed by a vowel and rarely in word-final position. Word-medially, it starts on average earlier and lasts longer than half of the fricative. Word-initially, it usually begins after the midpoint of the sound. While voiced AE trills are articulatorily possible, they do not occur in our data.

References

Arkhipov, A., Daniel, M., Belyaev, O., Moroz, G., Esling, J. H. 2019. A reinterpretation of lower-vocal-tract articulations in Caucasian languages. Proceedings of the 19th ICPhS, Melbourne, 1550–1554.

A. Arkhipov, M. Daniel, A. Shiryaev, E. Shepel. 2023. Evaluating Formant estimations and Discrete Cosine Transform to differentiate between pharyngeal fricatives in Mehweb. In: R. Skarnitzl & J. Volín (eds.), *Proceedings of the 20th ICPhS.* Guarant International. P. 3407–3411.

Esling, J. H. 2005. There are no back vowels: The laryngeal articulator model. *Canadian Journal of Linguistics*. 50, 13–44.

Esling, J. H., Moisik, S. R., Benner, A., Crevier-Buchman, L. 2019. *Voice Quality: The Laryngeal Articulator Model*. Cambridge: CUP.

Moroz, G. 2019. Phonology of Mehweb. In: Daniel, M., Dobrushina, N. & Ganenkov, D. (eds.), *The Mehweb language: Essays on phonology, morphology and syntax*. Language Science Press, 17–37. DOI:10.5281/zenodo.3402056

An investigation of syllable position /l/ allophony in L2 English learners using Word Error Rate as an index of phonetic proficiency

Anisia Popescu[1,2], Lori Lamel[1,2], Ioana Vasilescu[1,2], Laurence Devillers[1,2]

[1]LISN, CNRS
[2]Université Paris Saclay
anisia.popescu@universite-paris-saclay.fr

Introduction. The present study compares the production of L2 English lateral consonant allophony of L1 French and L1 Japanese speakers using the word error rate (WER) of an automatic speech recognition system (ASR) output to determine phonetic proficiency. English dialects exhibit a syllable position allophony which opposes clear(er) [l] in onsets and dark(er) [ɫ] in coda and syllabic positions. The acoustic differences between the two English /l/ allophones are well documented: clear /l/ has high F2 and low F1; dark /l/ has low F2 and higher F1 (Sproat & Fujimura, 1993). A staple measure of /l/ darkness is defined as the difference between the second and first formants (F2 – F1). The difference is larger in clear /l/ and lower in dark /l/ (Recasens, 2012). Neither French, nor Japanese distinguish between the two lateral allophones: French is known to have clear [l] in all positions and the Japanese phonological system does not include a /l/ phoneme but has another phoneme corresponding to both the rhotic and the lateral consonant, generally considered to be a tap /ɾ/. Only a few studies have looked at the phonetic detail of the acquisition of the lateral allophony in L2 English (L1 Japanese: Nagamine, 2022; L1 French and/or Spanish: Colantoni et al. 2023; King & Feragne 2017, Barlow, 2014) and to our knowledge they focused only on/ l/-words read in isolation and in controlled carrier sentences without including proficiency in their analysis. The present study investigates the acoustic characteristics of /l/ production in read texts taking phonetic proficiency levels into account. Low proficiency L2 English learners are expected to show small or no differences in F2-F1 values in onset vs. coda/syllabic /l/. High proficiency L2 English learners should exhibit differences in F2-F1 values between the two varieties of /l/. This pattern should hold independent on the participant's L1. Nagamine's (2022) study on Japanese speakers found differences in production between the two allophones but included only participants with high proficiency. Colantoni and colleagues (2023) tested speakers of different French dialects that had been living in Canada and did not include proficiency level in their analysis. Self-reported proficiency levels based on foreign language assessment are usually a measure of global proficiency that does not necessarily correlate with phonetic proficiency. Having access to native speaker judgments is not always straightforward and can be time-consuming. We therefore opted for an alternative method, which, like native speaker judgments, relies on the acoustic characteristics of global accent rather than an overall proficiency level of second language acquisition – ASR for English with acoustic models trained on native speech production. The advantage relying on acoustic models trained on large amounts of data is that the acoustic qualities of different segments are based on a set of homogeneous features that reflect statistical patterns found in actual language use.

Methods. Participants were 22 French (12 female) and 18 Japanese (11 male) native speakers. All participants were recorded reading the same three beginner-, intermediate- and advanced-level texts (see Kobylyanskaya etal., 2023). The acoustic signals of the productions were forced aligned using WebMAUS (with English US as a reference language). All lateral consonant tokens were hand-corrected in Praat (Boersma & Weenik, 2022). A total of 40 different words containing singleton /l/s (17 in onset, 19 in coda and 4 in syllabic position) were included in the analysis. Segmental duration and formant measures (F1, F2 and F3) at the midpoint of the lateral were extracted for all tokens. The difference between the second and first formant (F2-F1) was used to assess darkness in the lateral.

Participants were split into three proficiency level groups (beginner, intermediate, proficient) using a classification measure based on word-error-rate (WER) calculated using the WER() function in Matlab from the output of an unbiased (no text transcription was available) ASR system (Lamel etal., 2011). The WER classification was benchmarked against the self-reported proficiency levels, which were based on language assessment tests (IELTS, TOEIC, TOEFFL, CAE etc.). /l/ darkness (F2-F1) was analyzed using linear mixed effects models (*lme4* package Bates etal., 2015). *Syllable position* (onset, coda, syllabic), *language* (French, Japanese), *proficiency* (beginner, intermediate, proficient), *sex* (female, male), flanking *vowel position* (front, mid, back) and interaction terms between *syllable position* and *language*, *proficiency* and *vowel position* were used as predictors. *Participant* and *word* were included as random factors with intercepts. A second model with *duration* of the /l/ segment as a response variable was run using the same predictor and random effects structure.

Results. /l/ darkness results show that Japanese speakers produce overall darker /l/s than French speakers (Est. - 215Hz p-value < .001). This could be a language specific difference or a by-product of the larger proportion of male speakers in the Japanese participant pool. For French speakers proficiency level plays a significant role in lateral allophone distinction: proficient speakers produce darker /l/ in coda (Est. -285Hz, p-value <.0001) and syllabic (Est. -298Hz, p-value <.05) position. Intermediate speakers produce darker /l/ only in coda position (Est. -173Hz, p-value <.01) but not in syllabic position. Japanese speakers present a different pattern, with speakers classified by the WER measure as

proficient not significantly differing from beginner speakers. Japanese intermediate speakers however show significantly darker /l/s than beginners in both coda (Est. -136Hz, p-value <.01) and syllabic (Est. -201Hz, p-value<0.05) positions. Both French and Japanese speakers exhibit coarticulation with the flanking vowel. /l/ produced in the context of a back vowel, independent of syllable position, is produced as darker when compared to /l/ produced in the context of front vowels (French: Est. -177Hz, p-value < .05; Japanese: Est- -163Hz, p-value <.05). Interaction effects between vowel and syllable position are not significant for either French or Japanese speakers. As to be expected *Sex* plays a significant role with male participants producing overall lower formants for both Japanese and French speakers.
Duration results show that /l/ duration plays a role only in the case of proficient Japanese speakers which produce longer /l/s in coda (Est. 17ms, p-value < .01) and syllabic position (Est. 30ms, p-value <.01). French speakers do not differentiate lateral allophones based on duration.

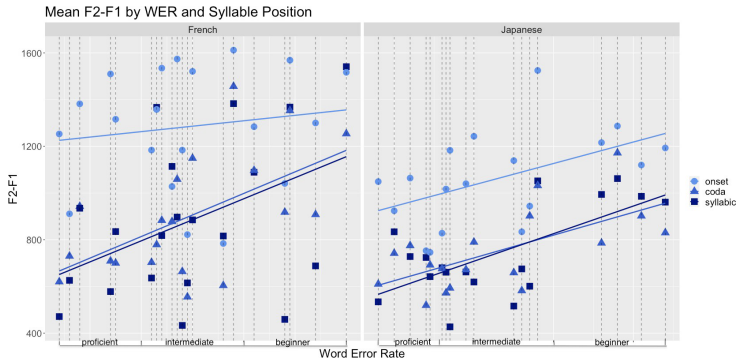


Figure 1: Mean F2-F1 measures per Word Error Rate values and Syllable position.
Fitted lines are provided for each syllable position level (onset, coda and syllabic /l/).

Discussion. The present study investigates the production of L2 English syllable position lateral consonant allophony in 40 speakers of L1 French and L1 Japanese. The acoustic analysis of the lateral allophones was related to the proficiency levels of the participants which was determined from the word error rate measured from the output of an automatic speech recognition system trained on native English data. Results show that while French speakers follow the predicted pattern of more proficient speakers producing more differentiated lateral allophones, the Japanese intermediate speakers produce the more differentiated allophones. Duration results also indicate a difference between French and Japanese speakers. In their study on /l/ allophony, Sproat and Fujimura (1993) found that /l/ is darker in longer rimes and suggested that the specific articulatory target of dark /l/ (i.e., tongue dorsum retraction) is more readily reached in longer rimes. This result was confirmed by Yuan and Liberman (2009) in a large corpus forced alignment study. The present study failed to find durational differences in French speakers but found durational effects for proficient Japanese speakers compared to beginner and intermediate speakers. This could be an indication that French and Japanese learners of L2 English tune into and adopt different phonetic strategies when producing phonetically different sounds in L2. The reported differences in the study do not appear when running the models using the self-reported proficiency levels, suggesting that using speech technology derived methodologies can provide more fine-grained phonetic classifications.

References

Barlow, J. (2014). Age of acquisition and allophony in Spanish-English bilinguals. *Frontiers in Psychology. 5. Article 288*

Bates, D., Maechler, M., Bolker B. & Walker, S. (2015) Fitting Linear Mixed-Effects Models Using lme3. *Journal of Statistical Software, 67(1), 1-48.*

Boersma, P. & Weenink, D. (2022). Praat: doing phonetics by computer from http://www.praat.org/.

Colantoni, L., Kochetov, A., & Steele, J. (2023). Articulatory Insights into the L2 Acquisition of English /l/ Allophony. *Language and Speech. 1-33.*

King, H. & Feragne, E. (2017). The effect of ultrasound and video feedback on the production and perception of English liquids by French learners. *Phonetics and Phonology in Europe. Cologne, Germany.*

Kobylyanskaya, S., Vasilescu, I., Augereau, O., Devillers, L. (2023) Vers la compréhension des difficultés de lecture en l2 à travers des paramètres acoustiques et de mouvement des yeux. 11ème Conférence sur les Environnements Informatiques pour l'Apprentissage Humain. EIAH2023

Lamel, L., etal. (2011). Speech recognition for machine translation in Quaero. In *Proceedings of the 8th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 121–128, San Francisco, California.

Moore, J., Shaw, J., Kawahara, S. & Arai, T. (2018). Articulation strategies for English liquids used by Japanese speakers. *Acoustical Science and Technology, 39*(2), 75-83.

Nagamine, T. (2022). Acquisition of allophonic variation in second language speech: An acoustic and articulatory study of English laterals by Japanese speakers. Proc. Interspeech 2022, 644-648.

Recasens, D. (2012). A cross-language acoustic study of initial and final allophones of /l/. *Speech Communication*, *54*(3), 368-383.

Sproat R. & Fujimura, O. (1993) Allophonic variation in Enlish /l/ and its implications for phonetic implementation. *Journal of Phonetics. 21(2), 291-311.*

Yuan, J. & Liberman, M. (2009). Investigation /l/ variation in English through Forced Alignment. Proc. Interspeech 2009, 2215

Acceleration peaks as representation of activation strength

Malin Svensson Lundmark

Lund University, Sweden; Queen Margaret University, UK
malin.svensson_lundmark@ling.lu.se

Introduction. Recent findings show that segment transitions can be accounted for by *acceleration peaks* (Svensson Lundmark 2023). Acceleration peaks occur when a mass changes its velocity the most, which it does in connection with changing direction (Eager et al. 2016), i.e. when an articulator either moves towards a speech posture or moves away from it. For example, for a speaker to shape a tongue tip constriction, the tip moves fast (a *velocity peak*) and then slows down rapidly (a *deceleration peak*; the force added is in the opposite direction of velocity). The tongue tip stays in position while forming a static speech posture, changes direction mid-posture, and then moves rapidly away again (a positive *acceleration peak*, followed by a *velocity peak*).
The present production study builds on previous research on the timing of acceleration peaks of various articulators (Svensson Lundmark 2023; Svensson Lundmark & Erickson 2024), and expands this line of research by investigating the magnitude of acceleration peaks in two discrete activation strengths: stressed and unstressed syllables. Investigating acceleration peaks in articulation of syllables is motivated by previous research on articulatory prosody. However challenging the study of the dynamic and highly complicated coordinated articulatory movements in the orofacial region is, previous findings have shown that articulatory movements (e.g., in the jaw) are faster, larger and longer in duration when syllables are prominent or stressed (de Jong 1995; Mücke & Grice 2014, van Summers 1987; Cho & Keating 2002; Erickson et al. 2012). This study is inspired by the concepts of *localized hyperarticulation* (de Jong 1995), and by *embodied phonetics* where focus is on groupings of muscles and nerves to move structures rather than discrete articulators (Gick 2019). Hence, this study assumes that articulation is different between stressed and unstressed syllables because of a difference in activation strength: a stressed syllable has overall more activation, more force added, than an unstressed syllable. More activation strength in stressed syllables would motivate the results on e.g. a lower jaw and faster movements as indicated by previous research. As a main representation of activation strength in the present study we are using acceleration peak magnitude at the release of an onset consonant.

Methods. Levels of activation strengths at onset release are achieved by comparing acceleration peak magnitude between a stressed syllable and an unstressed syllable. Velocity peak magnitude during the release is also included as a control feature, as this has previously been proven to vary dependent on syllable strength (de Jong 1995; Mücke & Grice 2014). Also included is timing of the acceleration peak, which is measured as the distance from zero-crossing velocity, as this is the point where the movement has changed direction and the articulator is moving in the opposite direction. Speech sounds are restricted to /m/ and /p/. Mouth cavity openness is investigated by including both lower lip and mandible lowering, as motivated by previous research on jaw-lip coordination showing tendencies of both synchronous and asynchronous kinematic behavior (Gracco 1994; Svensson Lundmark & Erickson 2023).
The method used is Electromagnetic articulography (EMA). The data is from an EMA corpus with 18 South Swedish speakers recorded at the Lund University Humanities Laboratory with the EMA system Carstens AG501 (250 Hz). Speakers read leading questions and target sentences (to elicit target words as given information) from a prompter, each set repeated eight times. In both target words (/mama/ and /papa/) the first syllable is stressed and the second unstressed. The present study uses data from EMA sensors on the lower lip (LL) and the lower incisors (mandible, JW) (see detailed information on the full corpus in Svensson Lundmark 2023). Post-processing of signals was done in Carstens software, after which the data were transferred to R for calculation and analysis. Both vertical and horizontal positions were used to calculate tangential velocity and acceleration. The acceleration signal has been filtered and smoothed using a low-pass filter, the R function *loess* (a time window of 0.02 seconds). Landmarks were collected semi-automatically in R: the author visually inspected each speaker's acceleration and velocity landmarks, and adjusted the time frames of the automatic script when justified. Welch Two Sample t-test was used to compare stressed and unstressed syllables, and Pearson correlation tests to assess the relationship between timing and magnitude.

Results. Preliminary findings (**Figure 1a**: 10 speakers, 165 tokens) on magnitude of acceleration peak show that the acceleration peak is higher in the stressed syllable both for LL ($t(297.32) = -5.3$, $p<.001$), and for JW ($t(283.74) = -2.8$, $p<.01$). Notably, there is overall more acceleration peak magnitude for LL than for JW in both the stressed syllable ($t(244.5) = 18.9$, $p<.001$) and the unstressed syllable ($t(255.52) = 19.3$, $p<.001$). Velocity peaks are also higher (=faster movements) in the stressed syllables than in the unstressed syllables, both for LL ($t(304.91) = -6.0$, $p<.001$), and for JW ($t(310.76) = -7.6$, $p<.001$). LL is moving faster than JW, in both the stressed syllable ($t(244.5) = 18.9$, $p<.001$) and the unstressed syllable ($t(253.83) = -18.8$, $p<.001$).

Figure 1b shows acceleration peak timing in the stressed syllable, which is the correlation between velocity peak magnitude (x-axis; how fast LL and JW are moving) and when in time the acceleration peaks occur (y-axis). There is no correlation, or a weak negative relationship, for a stressed LL (/m/: *r*=.04; /p/: *r*=-.23) and for a stressed JW (/m/: *r*=-.31; /p/: *r*=-.15). The unstressed syllables (no figure) display weak or moderate positive relationships, for an unstressed LL (/m/: *r*=.21; /p/: *r*=.24) and for an unstressed JW (/m/: *r*=.56; /p/: *r*=.13).

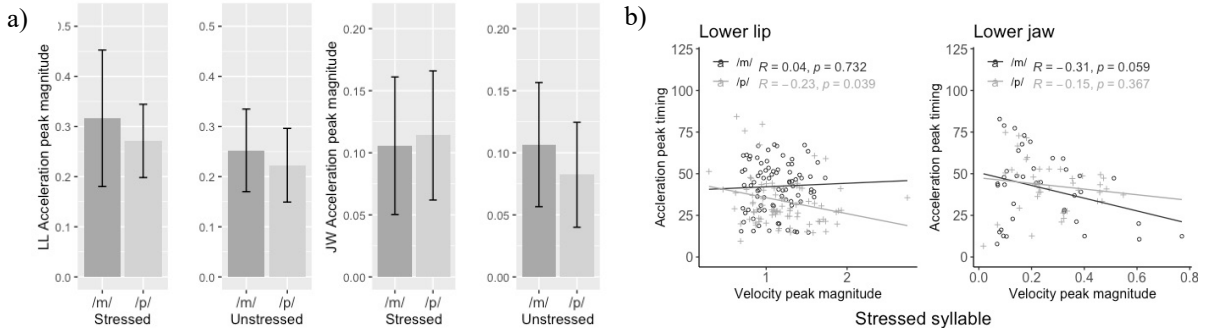


Figure 1: *Preliminary results on (a) acceleration peak magnitude in stressed and unstressed syllables; (b) acceleration peak timing in stressed syllables.*

Discussion. Lower lip and lower jaw both display higher acceleration peaks, as a representation of activation strength, and are also faster, in stressed syllables than in unstressed syllables: it appears that the amount of force added (acceleration peak magnitude) affects how fast the articulators move. In addition, the lips are moving faster than the jaw, which could be related to a lower mass (=move faster), but also due to more force added because of the lips' function as the primary articulator. The results could be interpreted as a stressed syllable is a result of a hyper-articulated lip constriction, challenging the notion of the jaw as the syllable articulator governing syllable strength (see e.g. Erickson et al. 2012). However, the relationship between the level of activation strength and the purpose of the movement needs more investigation. For example, how is acceleration peak magnitude related to the nature of the lip constrictions and intra-oral pressure in different stops?

Timing of the acceleration peaks are not related to how fast the articulators are moving, and by extension neither related to the size of the acceleration peaks. This seemingly uncoordinated phenomena might indicate that timing of the acceleration peak is functional in a different sense than the acceleration peak magnitude (the level of activation strength). These results could be interpreted as evidence of fixed time windows, and point towards a possible connection between timing of acceleration peaks and phase-locked neural oscillations and, as previous research has shown, segment transitions. The present study aims to stress the significance of acceleration peaks in speech, and to showcase the applicability of them in research on articulatory prosody.

References

Cho, T., & Keating, P. (2009). Effects of initial position versus prominence in English. J. Phon. 37(4), 466-485.

Eager, D., Pendrill, A.-M., and Reistad, N. (2016). "Beyond velocity and acceleration: Jerk, snap and higher derivatives," Eur. J. Phys. 37(6), 065008.

Erickson, D., Suemitsu, A., Shibuya, Y. and Tiede, M. (2012). "Metrical structure and production of English rhythm," Phonetica, 69(3), 180–190.

Gick, B. (2019). How bodies talk. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (eds.) Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia (pp. 20-24). Canberra, Australia: Australasian Speech Science and Technology Assoc. Inc. 2019.

Gracco V. L. (1994). Some organizational characteristics of speech movement control. J. Speech. Lang. Hear. Res. 37(1), 4–27. https://doi.org/10.1044/jshr.3701.04

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. J. Acoust. Soc. Am. 97(1), 491-504.

Mücke, D., & Grice, M. (2014). The effect of focus marking on supralaryngeal articulation–Is it mediated by accentuation? J. Phon. 44, 47-61.

Svensson Lundmark, M. (2023). Rapid movements at segment boundaries. J. Acoust. Soc. Am. 153 (3), 1452–1467. https://doi.org/10.1121/10.0017362

Svensson Lundmark, M., & Erickson, D. (2023) Comparing apples to oranges - asynchrony in jaw & lip articulation of syllables. In Proc. of the 20th International Congress of Phonetic Sciences, Prague, Czech Republic.

Svensson Lundmark, M., & Erickson, D. (2024) Segmental and syllabic articulations: a descriptive approach. J. Speech. Lang. Hear. Res. https://doi.org/10.1044/2024_JSLHR-23-00092

Van Summers, W. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. J. Acoust. Soc. Am. 82(3), 847-863.

Are long and short vowels articulatorily different? Spatial and durational effects of vowel length in Thai

Sireemas Maspong[1], Francesco Burroni[1]

[1]Institute for Phonetics and Speech Processing, LMU Munich

s.maspong@phonetik.uni-muenchen.de, francesco.burroni@phonetik.uni-muenchen.de

Introduction. In Thai, vowel length distinguishes all monophthongs. Previous research has demonstrated that duration primarily cues the length contrast for all vowel pairs acoustically and perceptually (Abramson, 1964; Abramson & Ren, 1990). However, studies have also indicated marginal differences in quality between short and long vowels in Thai, with short vowels tending to be more centralized in the acoustic vowel space (Abramson, 1964). Despite these findings, our understanding of the articulatory characteristics underpinning vowel length contrast in Thai, particularly concerning spatial and temporal aspects, remains limited. This paper presents an investigation into the articulatory features of long and short vowels in Thai, drawing upon results obtained from electromagnetic articulography (EMA). Our study reveals distinct spatial and temporal characteristics associated with the production of long vowels, contrasting with their short counterparts. These results challenge the conventional assumption regarding Thai vowel length, which typically views long vowels as short vowels with longer durations, thus questioning the undershooting account proposed by Lindblom (1963).

Methods 1. Articulatory data were collected from 7 native Thai speakers using a 3D AG501 Carsten Electromagnetic Articulography (EMA). The participants were instructed to produce nonce words in the format mVm, containing either /a(ː)/ or /i(ː)/ vowels, with variations in Mid, Low, or High tones. To optimize tongue vertical movement and facilitate landmarking, target words were embedded in distinct carrier sentences. Specifically, words with /a(ː)/ were surrounded by words with /iː/ vowel, while words with /i(ː)/ were surrounded by words with /aː/ vowel. The choice of bilabial onset and coda was deliberate to minimize conflicting demands on tongue movement. Participants were directed to execute the stimuli at three different speech rates, introducing variability in vowel duration. Vocalic gestures were identified by tracking tongue movement for high vowels (/i(ː)/) and jaw movements for low vowels (/a(ː)/). Landmarking of tongue movement was executed on the first principal component (following the approach of Sorensen & Gafos, 2015) derived from three dimensions of Tongue Tip, Tongue Body, and Tongue Back sensors. Conversely, landmarking of jaw movement was conducted based on the vertical displacement of the jaw. From the landmarking process, five measurements were extracted for each articulatory trajectory: (i) Minimum Jaw Height for vowel /a(ː)/ and Maximum Tongue Body Height for /i(ː)/, (ii) duration of the articulatory steady state (the duration between articulatory target and release landmarks), (iii) movement amplitude, (iv) peak velocity from onset to target, and (v) stiffness (calculated as the ratio of peak velocity to movement amplitude). Linear mixed-effect regressions were fit, treating each measurement as a dependent variable. Fixed effects included vowel length (long or short), utterance duration (z-scored and normalized by subtracting the target vowel duration, following the methodology of Tilsen & Tiede, 2023), and their interaction. Additionally, subject was included as a random intercept in the analysis.

Results 1. Preliminary results indicate distinctive articulatory patterns associated with long vowels, characterized by a statistically significant lower Minimum Jaw Height for /aː/ and a higher Maximum Tongue Height for /iː/. Long vowels exhibit a significantly longer duration at the steady state, a broader movement amplitude, a faster peak velocity, and a lower stiffness when compared to their short counterparts. **Figure 1**, depicting jaw movement trajectories of short and long /a(ː)/ with identical durations, highlights the consistently lower jaw position of long vowels from gestural onset throughout the trajectories.

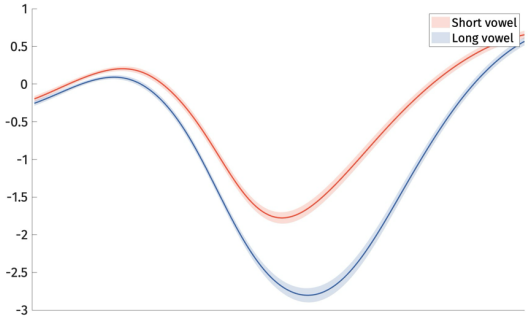


Figure 1: *Trajectories of Jaw Height of short /a/ and long /aː/ vowels with -1 to 1 z-scored duration*

Moreover, our analysis reveals a significant impact of utterance duration, reflecting speech rate, on all five measured articulatory parameters, regardless of vowel length. Notably, the effect size of utterance duration is more pronounced for long vowels than for short vowels, as we observed a significant effect of the interaction between vowel length and utterance duration. Both short and long vowels exhibit an increased duration of the articulatory steady state as the utterance duration increases (indicating a slower speech rate); however, short vowels show a comparatively smaller increase in duration compared to long vowels.

Methods 2. After observing that long vowels exhibit lower Minimum Jaw Height for /a(ː)/ and a higher Maximum Tongue Height for /i(ː)/ compared to short vowels, we proceeded to test Lindblom's (1963) undershoot account. This account proposes that short vowels are more centralized than long vowels due to the limited duration of short vowels, which restricts the time to reach the target and results in undershooting of the vowel target. The prediction is that if the spatial difference across vowel length arises from the duration difference of the short and long vowels, we would not observe differences in jaw or tongue height if short and long vowels have equal time to reach the target. To test this prediction, we extracted an additional measurement from the same dataset: duration to target (the duration from vowel articulatory onset to vowel articulatory target). We employed linear mixed-effect regressions, treating Minimum Jaw Height for /aː/ and Maximum Tongue Height for /iː/ as dependent variables. Fixed effects included vowel length (long or short), duration to target, and their interaction. Subject was once again included as a random intercept in the analysis.

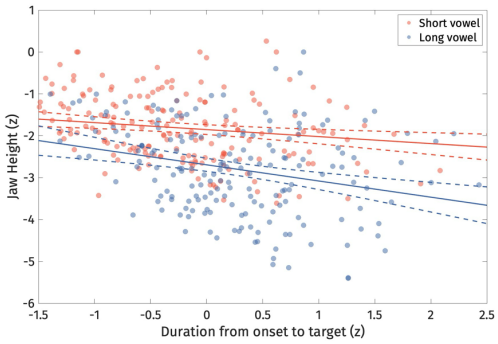


Figure 2: *Normalized Jaw height of short /a/ and long /aː/ vowels as a function of normalized time to target*

Result 2. Preliminary results indicate that even when short and long vowels have equal time to reach the target, their Minimum Jaw Height and Maximum Tongue Height exhibit significant differences, as shown in **Figure 2**. Time to target does not affect Minimum Jaw Height and Maximum Tongue Height for short vowels, as we did not observe a significant effect of the time to target fixed effect. However, time to target negatively correlates with Minimum Jaw Height for long /aː/ and positively correlates with Maximum Tongue Height for long /iː/, as the interaction of vowel length and time to target was significant.

Discussion. Our articulatory investigation reveals not only durational differences but also distinct spatial features characterizing long and short vowels in Thai. While our results initially appear compatible with the undershoot account, the combined observed differences in peak velocity and stiffness between long and short vowels may not entirely align with this interpretation. Particularly, differences in stiffness seem to be category-specific, consistent with previous findings on singleton vs. geminate stops (Löfqvist, 2005). Furthermore, even when short and long vowels have an equal time to reach the target, they still exhibit distinct spatial features, namely minimum jaw height or maximum tongue height. These findings resonate with articulatory studies of vowel length in other languages, such as Australian English (Ratko et al., 2023), suggesting that the distinction between long and short vowels in Thai extends beyond a simple duration-related difference and a difference in target. Instead, it points towards unique articulatory characteristics and control regimes associated with each vowel length.

References

Abramson, A. S. (1962). The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments. *International Journal of American Linguistics*, 28(2). Bloomington: Indiana University.

Abramson, A. S., & Ren, N. (1990). Distinctive vowel length: duration *vs.*spectrum in Thai. *J Phon*. 18(2), 79-92.

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *JASA*, *35*(11), 1773–1781.

Löfqvist, A. (2005). Lip kinematics in long and short stop and fricative consonants. *JASA*, 117(2), 858–878.

Ratko, L., Proctor, M., & Cox, F. (2023). Articulation of vowel length contrasts in Australian English. *JIPA*, *53*(3), 774–803.

Sorensen, T., & Gafos, A. I. (2015). Changes in vowel velocity profile with vowel-consonant overlap. *ICPhS*.

Tilsen, S., & Tiede, M. (2023). Parameters of unit-based measures of speech rate. *Speech Communication*, 150, 73–97.

Enhancing lip contrasts between /u/ and /y/ in Taiwan Mandarin

Chenhao Chiu[1,2,3], Cheng-Hsiang Chang[4], Jian-Zhi Huang[1], Po-Hsuan Huang[1]

[1] Graduate Institute of Linguistics, National Taiwan University
[2] Graduate Institute of Brain and Mind Sciences, National Taiwan University
[3] Neurobiology and Cognitive Science Center, National Taiwan University
[4] Department of Economics, National Taiwan University

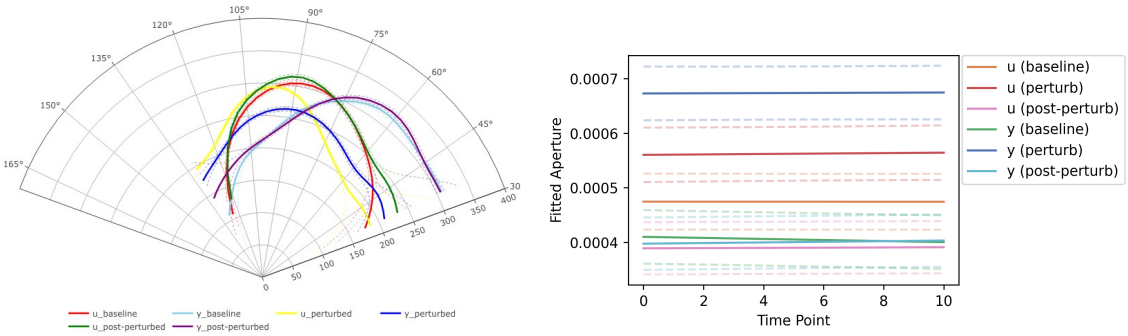
chenhaochiu@ntu.edu.tw, felixchang0423@gmail.com, r10142007@ntu.edu.tw, benson32169@gmail.com

Introduction. Mandarin distinguishes three high vowel phonemes /i, u, y/ in terms of lip postures and tongue backness (Lin, 2007). While lip postures provide more visual cues than tongue backness (Lisker & Rossi, 1992), articulatory contrasts in lips between /u/ and /y/ appear to be rather speaker-dependent (Chiu and Huang, 2023). Despite both being rounded vowels with the same feature (i.e., [+labial]), there are subtle differences in lip postures: /u/ requires pulling of the corners of the lips and more protrusion (*endolabial*, according to Catford's 1988 terminology), whereas /y/ involves upper and lower lip approximation with less protrusion (*exolabial*; Catford, 1988). However, these subtle differences do not constitute phonemic contrasts, rendering lip postures a secondary articulation compared to tongue position (i.e., backness). The role of such secondary articulatory mechanisms and their indispensability have been less discussed. One way to approach this is through perturbation design in speech production. Perturbation studies since the 1980s have explored a wide range of issues, including feedback and feedforward mechanisms (Abbs and Gracco, 1984), motor equivalence (Perkell et al., 1993), among others. Perhaps one of the most critical aspects in perturbation studies is the compensatory behavior in response to perturbation. For instance, perturbation to the lower lip during the production of a bilabial sound can induce compensatory movement from the upper lip to achieve the intended bilabial closure (Gracco and Abbs, 1986; Shaiman and Gracco, 2002). These compensations from other articulatory effectors validate their own role and importance in order to achieve the phonemic target. In other words, successful speech production may involve both primary and secondary articulation. However, when one of these is compromised, the role and importance of the other articulatory mechanism can be enhanced. Following this logic, this study investigates the enhancement of lip postures, which are the secondary articulation for high rounded vowels as in Mandarin, through the design of perturbation. The prediction is that lip contrasts, such as endolabial vs. exolabial, would be enhanced when the primary articulation is perturbed.

Methods. Three native speakers (1 female) of Taiwan Mandarin participated in the study. None of them reported any visual or hearing impairment. The main task involves perturbation to the tongue position by keeping a bolus (i.e., a hard-boiled egg) in the mouth during phonation. Participants were instructed to produce /u/ and /y/ with 20 repetitions (self-paced) while making audible contrasts between /u/ and /y/. Ultrasonography was employed to record tongue movements (in .mp4 at 60 fps) during phonation, simultaneously with acoustics recorded in .wav at 44.1K Hz. Tongue postures at the midpoint of the acoustic recordings were obtained and included for statistical analyses. Baseline tongue positions before bolus perturbation and post-perturbation tongue positions were also acquired. To monitor the changes of lip postures, a high-speed camera was positioned in front of the participants at approximately a distance of 80 cm, recording at 120 fps. Markers on the nose (as reference), upper vermillion, lower vermillion, and the rightmost corner of the mouth were placed to quantify the degree of lip protrusion. Videos processing was conducted using MediaPipe (Google Inc.) to measure lip aperture. Individual measurements, including tongue posture, upper lip protrusion, lower lip protrusion, lip aperture, and F1 – F3, were analyzed using Generalized Additive Mixed Models (GAMMs; Wieling, 2018).

Results. When perturbed by an egg bolus, the tongue tip and blade were substantially depressed, leading to a larger proportion of the tongue moving posteriorly due to muscular hydrostatics. As a result, the postural contrasts between /u/ and /y/ diminished (as shown in Figure 1a, with yellow and blue lines representing each vowel, respectively), compared with baseline and post-perturbation conditions. This confirms successful perturbation of tongue position during the production of /u/ and /y/. In general, more upper and lower lip protrusion was observed in the production of /u/ than /y/ without perturbation. However, when perturbed, both lips became more protruded, yet the degree of protrusion remained contrastive, being more pronounced for /u/ than /y/ (not shown in this abstract). Regarding lip aperture, there was no observable difference between /u/ and /y/ at baseline. However, the lip aperture was larger for /u/ than /y/ in the baseline, though the difference was not statistically significant (as indicated in Figure 1b, with orange and green lines for each vowel, respectively). Upon perturbation with the egg bolus, there was an increase in lip aperture for both vowels, and crucially, the contrast in aperture between /u/ and /y/ became statistically significant (Figure 1b, depicted with red and

blue lines, respectively). Acoustic formants were also examined to confirm that the vowels were produced as instructed. Our findings indicate that the perturbation had no effect on F1. When perturbed, F2 for /y/ significantly lowered, reflecting the depressed tongue tip and blade caused by the bolus, resulting in more tongue volume towards the back of the mouth. However, for /u/, F2 showed no significant change between baseline and perturbation conditions. Additionally, baseline /y/ was associated with intrinsically lower F3 than baseline /u/. For /y/, F3 lowered when perturbed, suggesting a more pronounced lip posture in response to tongue perturbation. In contrast, perturbation had no effect on F3 for /u/. Notably, both /u/ and /y/ demonstrated after-effects in lip postures (i.e., lowered F3) but not in tongue positions, as depicted in Figure 1a.



(a) (b)

Figure 1: (a) *Tongue postures for /u/ and /y/ across conditions. Tongue tip on the right.* (b) *Lip aperture results across conditions over 10 time points.*

Discussion. The present study investigates whether lip contrasts between the high rounded vowels /u/ and /y/ can be enhanced when the primary articulatory contrasts of these vowels (i.e., the tongue) are perturbed. Although lip postures may be considered secondary or non-phonemic in everyday speech, our findings suggest that these secondary articulatory postures can become indispensable and even enhanced. During bolus perturbation, the tongue posture was significantly constrained, resulting in limited postural contrasts. Despite this, the tongue root continued to advance in the production of /y/, and remained elevated for the target /u/, indicating that the tongue still strives to reach its target position and posture. Given the limited degrees of freedom of the tongue under these conditions, the acoustic contrasts between /u/ and /y/ could only be achieved through other articulatory gestures – namely, the lips. A larger aperture was noted in the perturbed productions compared to their baseline counterparts. More crucially, larger aperture associated with /y/ than with /u/ remain consistent in both baseline and perturbed productions. In summary, this study underscores the adaptive flexibility of articulatory mechanisms, demonstrating that even when primary articulators are constrained, secondary articulatory features, such as lip postures, can compensate effectively to maintain phonetic distinctions.

References

Abbs, J. H., & Gracco, V. L. (1984). Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech. *Journal of Neurophysiology, 51*(4), 705-723.

Catford, J. C. (1988). *A Practical Introduction to Phonetics*. Oxford: Clarendon Press.

Chiu, C., and Huang, P.-H. (2023). Lip postures of high vowels in Taiwan Mandarin. *Proceedings of the 2023 International Congress of Phonetic Sciences*, 1052 – 1056.

Gracco, V. L., & Abbs, J. H. (1986). Variant and invariant characteristics of speech movements. *Experimental Brain Research*, *65*, 156-166

Lin, Y. H. (2007). *The Sounds of Chinese.* Cambridge University Press.

Lisker, L., & Rossi, M. (1992). Auditory and visual cueing of the [± rounded] feature of vowels. *Language and Speech, 35*(4), 391-417.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., & Jordan, M. I. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel/u: A pilot ‘‘motor equivalence’’study. *The Journal of the Acoustical Society of America, 93*(5), 2948-2961.

Shaiman, S., & Gracco, V. L. (2002). Task-specific sensorimotor interactions in speech production. *Experimental Brain Research, 146*, 411-418.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics, 70*, 86-116.

Sub-visemic discrimination and the effect of visual resemblance on silent lip-reading

Maëva Michon[1,3], Pablo Billeke[2], Gabriela Galgani[3], Francisco Aboitiz[3]

[1] Praxiling Laboratory, UMR 5267, CNRS, Université Paul Valéry, Montpellier, France
[2] Laboratorio de Neurociencia Social y Neuromodulación, Centro de Investigación en Complejidad Social (neuroCICS), Facultad de Gobierno, Universidad del Desarrollo
[3] Laboratorio de Neurociencia Cognitiva y Evolutiva, Centro Interdisciplanrio de Neurociencia NeuroUC, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile

maeva.michon@hotmail.fr

Introduction.

Although speech perception has traditionally focused on auditory speech processing, there is now increasing evidence supporting the importance of visual speech perception (Lalonde & Werner, 2019; O'Sullivan et al., 2021). During face-to-face interactions, the movements of the speakers' mouths represent valuable information that optimizes listeners' comprehension. However, there is an ongoing debate about how phonemes and their visual counterparts, visemes are mapped (Bear & Harvey, 2017). An influential hypothesis claims that several phonemes can be mapped into a single visemic category (many-to-one phoneme-viseme mapping). In contrast, recent findings have challenged this view, reporting evidence for sub-visemic syllable discrimination (Files et al., 2015). The purpose of this study is to investigate whether Spanish words from the same lexical equivalence class are visually discriminable. The concept of lexical equivalence class refers to words that sound differently but look very similar on speakers' lips. For example, according to computational classifications phonemes /b/ and /m/ are thought to belong to the same phonemic equivalence class; as such the words "beet" and "meet" are from the same lexical equivalence class (Auer & Bernstein, 1995; Bernstein, 2012).

Methods.

To assess the effect of visemic information on lip-reading performance, we designed a forced choice word identification task in which participants had to identify target words displayed in silent video clips among 3 orthographic distractors differing in their visemic resemblance to target words. For instance, in Spanish, the distractor "pata" (SameVis) and the target word "bata" looks very similar since /b/ and /p/ are phonemes from the same visemic category, whereas the word "gata" (DifVis) is less resembling the target word since phonemes /g/ and /b/ belong to different visemic category. Finally, the distractor "bota" (DifVow) resembles the target word "bata" even less since it contrasts vowels instead of consonants in a different location in the distractor set (second letter) as compared to target. We hypothesize that finer-grained spatial information from mouth shape and orofacial articulatory sequences than previously thought is used during lip-reading. We predict that words from the same lexical equivalence class should be discriminable, allowing the perceiver to identify the target word above chance. Moreover, we expect that the error rate will be greater for distractors that belong to the same lexical equivalence class compared to those that do not.

Results.

We tested lip-reading performance against chance level and found that only the rate of Target responses was above chance level, with no difference for SameVis (at chance) and significant below chance levels for DifVis and Vow. In line with our hypothesis, the response rate for target words was significantly greater than responses for the distractor belonging to the same visemic category. Together, these results indicate that participants' performance on the lip-reading task was above chance and that they were able to extract sub-visemic variation in visual speech cues. Moreover, as expected, the mean error rate for distractors significantly decreased with decreasing visual resemblance from the target (Same Viseme > Different Viseme > Different Vowel), confirming that within the lexical equivalence class words are more challenging to identify than between lexical equivalence class words.

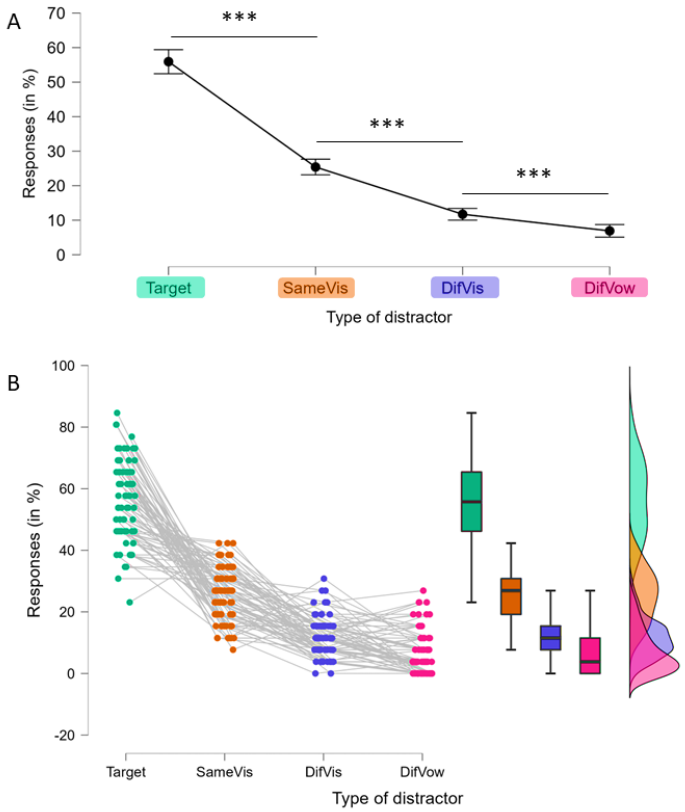


Figure 1: *a) Mean percentage of responses for target words and the different distractors. Error bars correspond to a confidence interval of 95%. (b) Raincloud plot showing the mean percentage of accuracy for each participant for each word type.*

Discussion.

Does /b/, /p/, and /m/ form a unique visemic category? Is there such a thing as a visemic category? Altogether, the results of the present study indicate that in Spanish it is possible to discriminate words within (what are thought to be) visemic categories, challenging the many-to-one phoneme-viseme mapping hypothesis. It could be the case that very subtle differences exist in orofacial positioning and/or in the timing of orofacial movement sequences that can be visually detected for discrimination. Further research is needed to reach a deeper understanding of which exact spatial and temporal features are available to human eyes to achieve fine-grained, within-viseme discrimination through silent lip-reading. Greater insights on this question could have important implications in the field of vision computer science to improve the development of applications or software for automatic lip-reading that could be helpful for patients who can move their lips and tongue but are not able to properly produce speech (after intubation and laryngeal cancer for instance).

References

Auer, E. T., Jr., & Bernstein, L. E. (1995). Lexical distinctiveness in lipreading: Effects of phonemic equivalence classes on the structure of the lexicon. The Journal of the Acoustical Society of America, 97(5_Supplement), 3361-3362. https://doi.org/10.1121/1.412683

Bear, H. L., & Harvey, R. (2017). Phoneme-to-viseme mappings: The good, the bad, and the ugly. Speech Communication, 95, 40-67. https://doi.org/10.1016/j.specom.2017.07.001

Bernstein, L. E. (2012). Visual speech perception. En E. Vatikiotis-Bateson, G. Bailly, & P. Perrier (Eds.), Audiovisual Speech Processing (pp. 21-39). Cambridge University Press. https://doi.org/10.1017/CBO9780511843891.004

Files, B. T., Tjan, B. S., Jiang, J., & Bernstein, L. E. (2015). Visual speech discrimination and identification of natural and synthetic consonant stimuli. Frontiers in Psychology, 6. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00878

Lalonde, K., & Werner, L. A. (2019). Infants and Adults Use Visual Cues to Improve Detection and Discrimination of Speech in Noise. Journal of Speech, Language, and Hearing Research: JSLHR, 62(10), 3860-3875. https://doi.org/10.1044/2019_JSLHR-H-19-0106

O'Sullivan, A. E., Crosse, M. J., Liberto, G. M. D., de Cheveigné, A., & Lalor, E. C. (2021). Neurophysiological Indices of Audiovisual Speech Processing Reveal a Hierarchy of Multisensory Integration Effects. The Journal of Neuroscience, 41(23), 4991-5003. https://doi.org/10.1523/JNEUROSCI.0906-20.2021

Acquiring tongue shape complexity in Scottish Gaelic consonants

Claire Nance, Sam Kirkham

Lancaster University

c.nance@lancaster.ac.uk, s.kirkham@lancaster.ac.uk

Introduction. This poster presents a pilot study of children's acquisition of complex consonant articulation. Cross-linguistic surveys indicate that children typically acquire consonants requiring fine motor control or multiple lingual gestures, such as affricates and rhotics, up to age 6-7 years (McLeod and Crowe 2018). This is due to developing control of lingual differentiation i.e. mastery of different tongue gestures and their coordination in time (Abakarova, Fuchs, and Noiray 2022). At the same time, minority language bilingual acquisition of sounds and structures not shared with the societally dominant language can be more protracted than in monolingual first language acquisition (Kennard 2018; Nance 2020).

Here, we investigate Scottish Gaelic-English bilingual children's production of consonants with secondary articulations in Gaelic. Specifically, we consider tongue shape complexity measured by the curvature of the midsagittal tongue shape (Dawson, Tiede, and Whalen 2015; Dokovova et al. 2023; Kabakoff et al. 2023). Scottish Gaelic has a system of palatalisation contrasts across the consonants (Nance and Ó Maolalaigh 2021). We focus on word-initial /l̪ʲ l̪ˠ n̪ʲ n̪ˠ/ in child and adult speech. These consonants are described as dental and palatalised/velarised and are therefore hypothesised to be produced with complex tongue shapes. We investigate the following questions: 1) Can measures of tongue curvature be extended to analysis of phonemic secondary articulations? 2) Are there age-related differences in tongue shape complexity? 3) Does complexity vary according to bilingualism background?

Methods. Data were collected from 22 Gaelic-English bilinguals on the Isle of Lewis, Scotland. Participants include 14 children in a Gaelic Medium primary school (a form of immersion schooling) aged 6-11 (3f, 11m); and 8 adults who work as Gaelic language professionals in Lewis and represent community target norms for the children (aged 21-72; 4f, 4m). Synchronous audio and ultrasound data were recorded in AAA with a headset microphone and a Telemed MicrUs ultrasound machine at approximately 92Hz frame rate (Articulate Instruments 2022). Participants wore a plastic helmet to hold the probe. Here, we analyse eight words containing word-initial sonorants, two repetitions of each (352 tokens in total).

Audio data were extracted from AAA, and the sonorant and following vowel segmented in Praat. TextGrids were uploaded back into AAA and tongue splines fitted to the sonorant-vowel interval using the DeepLabCut plugin (Wrench and Balch-Tomes 2022). We extracted the coordinates of the tongue spline at sonorant midpoint for analysis, rotated to the midsagittal plane of each participant. Tongue complexity was measured with Maximum Curvature Index using the Python script from Dawson, Tiede, and Whalen (2015). Data were analysed using linear mixed effects regression including whether the participant was an adult or a child, and their gender, as fixed effects. Nasals and laterals were analysed separately. Due to the small size of this pilot dataset, we provide qualitative observations on age-related differences within the children, as well as home language background, rather than statistical testing.

Results. Results indicate that children use less complex tongue shapes than adults in our dataset. Qualitative examination of the data indicates that children aged 6-8 in particular use less complex tongue shapes. Children aged 10-11 approach adult productions. The differences between children and adults are larger in lateral production than in nasals. In laterals, children with some Gaelic input at home produced more complex tongue shapes. There were few or no differences for home language background in nasals.

Discussion. Our analysis so far indicates that Dawson et al.'s method is also appropriate for phonemic secondary articula-

tions. Similar to previous work we find age-related differences in tongue complexity (Dokovova et al. 2023), tongue shape is least complex in 6-8 year olds. In the Gaelic context, acquisition of tongue shape complexity seems to be mediated by home language background to some extent suggesting an interaction of developmental and bilingualism factors. These results add an articulatory dimension to current understanding of input and phonological acquisition in minority language bilingual settings (Munro et al. 2005; Nance 2020).

References.

Abakarova, Dzhuma, Susanne Fuchs, and Aude Noiray (2022). "Developmental Changes in Coarticulation Degree Relate to Differences in Articulatory Patterns: An Empirically Grounded Modeling Approach". In: *Journal of Speech, Language, and Hearing Research* 65.9, pp. 3276–3299. DOI: 10.1044/2022_jslhr-21-00212.

Articulate Instruments (2022). *Articulate Assistant Advanced version 2.20.2*. Edinburgh: Articulate Instruments.

Dawson, Katherine, Mark Tiede, and D. H. Whalen (2015). "Methods for quantifying tongue shape and complexity using ultrasound imaging". In: *Clinical Linguistics and Phonetics* 30.3-5, pp. 328–344. DOI: 10.3109/02699206.2015.1099164.

Dokovova, Marie, Ellie Sugden, Gemma Cartney, Sonja Schaeffler, and Joanne Cleland (2023). "Tongue Shape Complexity in Children With and Without Speech Sound Disorders". In: *Journal of Speech, Language, and Hearing Research* 66.7, pp. 2164–2183. DOI: 10.1044/2023_jslhr-22-00472.

Kabakoff, Heather, Sam Pearl Beames, Mark Tiede, D. H. Whalen, Jonathan L. Preston, and Tara McAllister (2023). "Comparing metrics for quantification of children's tongue shape complexity using ultrasound imaging". In: *Clinical Linguistics and Phonetics* 37.2, pp. 169–195. DOI: 10.1080/02699206.2022.2039300.

Kennard, Holly (2018). "Verbal lenition among young speakers of Breton: Acquisition and maintenance". In: *New speakers of minority languages: Linguistic ideologies and practices*. Ed. by Cassie Smith-Christmas, Noel Ó Murchadha, Michael Hornsby, and Máiréad Moriarty. London: Palgrave Macmillan, pp. 231–249.

McLeod, Sharynne and Kathryn Crowe (2018). "Children's Consonant Acquisition in 27 Languages: A Cross-Linguistic Review". In: *American Journal of Speech-Language Pathology* 27.4, pp. 1546–1571. DOI: 10.1044/2018_ajslp-17-0100.

Munro, Siân, Martin Ball, Nicole Muller, Martin Duckworth, and Fiona Lyddy (2005). "Phonological acquisition in Welsh – English bilingual children". In: *Journal of Multilingual Communication Disorders* 3.1, pp. 24 –49.

Nance, Claire (2020). "Bilingual language exposure and the peer group: Acquiring phonetics and phonology in Gaelic Medium Education". In: *International Journal of Bilingualism* 24.2, pp. 360–375.

Nance, Claire and Roibeard Ó Maolalaigh (2021). "Scottish Gaelic". In: *Journal of the International Phonetic Association* 51.2, pp. 261–275.

Wrench, Alan and Jonathan Balch-Tomes (2022). "Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut". In: *Sensors* 22.3, p. 1133. DOI: 10.3390/s22031133.

Effects of fundamental frequency and spectral manipulations on speech production under delayed auditory feedback

Yasufumi Uezu, Masato Akagi, and Masashi Unoki

Japan Advanced Institute of Science and Technology, Japan

{y-uezu, akagi, unoki}@jaist.ac.jp

Introduction. Auditory monitoring of one's own speech sounds in real time plays an important role in the speech motor control. One experimental paradigm to elucidate the mechanisms of auditory-speech motor control is the delayed auditory feedback (DAF). When speaking under DAF, the produced speech is time-delayed and fed back to the speaker's auditory system. Speaking under DAF results in reduced speech fluency and increased speech duration. In particular, DAF with a delay of 200 ms is known to have an strong effect on speech (Lee (1950)). Typically, most experiments with DAF do not manipulate any information other than the delay. The feedback speech under such DAF is assumed to contain enough of the speaker's own acoustic individuality. Therefore, it is possible that under such DAF, the speaker's own speech is perceived as delayed, resulting in a decrease in speech fluency. Do differences in the acoustic individuality in the feedback speech affect speaking under DAF? Toyomura and Omori (2005) conducted speech experiments under DAF with the condition that the fundamental frequency (f_o) of the feedback speech is consistently shifted upward. The results showed that the larger the amount of shift in f_o given to the feedback speech, the smaller the DAF effect on speech. The acoustic individuality of speech can be altered not only by manipulating the f_o of speech, but also by stretching or contracting the spectrum of speech in the frequency direction. Therefore, it is possible to investigate the relationship between the acoustic individuality of feedback speech and the effect of DAF on speech by manipulating both the f_o and the spectrum of feedback speech to reduce speaker's individuality. In this study, we investigated the effects of differences in the acoustic individuality of feedback speech and the amount of delay on speaking under DAF through the experiments in which both the f_o and the spectrum of speech in the feedback were manipulated.

Methods. The experiment was conducted in a soundproof room. Participants were twelve adult male native Japanese speakers (mean 25.0 $\pm$ 1.8 years old). Uttered speech was recorded using a headset microphone (DPA d:fine, DAD6001). The speech was then amplified in a first audio interface (Focusrite Clarett+ 4Pre) and input to a laptop PC (Lenovo Thinkpad X1 carbon) via a second audio interface (Roland QUAD-CAPTURE). The input signal was delayed with the Audapter (Cai et al. (2008)) running on the laptop PC and then output. The output signal was manipulated by a vocal effector (Roland VT-4) and presented to the speaker as feedback speech via a mixer (YAMAHA MG10XUF) and headphones (Sennheiser 280mk2). The speech sound and the feedback speech sound were input to another laptop PC via the first audio interface and recorded as a stereo wav file in Audacity at $Fs = 48$ kHz, 16 bit. Prior to the experiment, the input and output gains were adjusted so that the sound pressure level (SPL) of the headphone output was 85 dB when the microphone input SPL was 94 dB. For the reading aloud task, two Japanese sentences of 21 mora each were selected, which have different difficulties to be read aloud in a normal listening environment. In this study, the amount of delay for the DAF was set to 0 ms, 100 ms, and 200 ms as the DELAY condition. The manipulation of the f_o and the spectrum of the feedback speech was achieved by controlling PIT and FMT parameters that can be manipulated via the vocal effector. The PIT conditions were set to 0, Up, and Down. These manipulations were equivalent to changing the f_o of the feedback speech to F_o, $F_o \times 1.8$, and $F_o/1.8$ [Hz], respectively, when the uttered speech f_o is F_o. The FMT conditions were set to 0, Up, and Down. These manipulations were performed such that the spectral center of gravity (CoG) of the feedback speech was changed to C, $C \times 1.8$, and $C/1.8$ [Hz], respectively, when the CoG of the uttered speech was C. First, the participants practiced reading Japanese sentences aloud. All participants confirmed that they could read all sentences aloud without any problems. After the practice, the main reading aloud task was performed with the headset microphone and headphones attached. The reading aloud task consisted of repeating the sequence as follows: One of the Japanese sentences was shown on the display for six seconds. In the main trial, the speaker was instructed to read the sentence aloud only once within 12 seconds. After the main trial, the text of the five Japanese vowels for the dummy trial appeared for four seconds. In the dummy trial, the speaker was asked to repeat the text three times within eight seconds. A total of 54 main trials (two DELAY $\times$ three PIT $\times$ three FMT $\times$ two sentences) were performed, divided into three blocks. The order of the main trials was randomized and counterbalanced across participants. For all 648 utterances obtained in the experiment, the start and end points of the speech sound were determined by waveform checking and listening

to the speech data, and the time duration between these points was used as the speech duration. A three-factor repeated measures analysis of variance (RM-ANOVA) was performed on the speech duration, with the degrees of freedom adjusted by Chi-Muller's ϵ when the assumption was rejected by Mauchly's sphericity test. Shaffer's test was used for multiple comparisons. Additionally, Dunnet's tests were performed on the speech duration under each DELAY with the [PIT, FMT] = [0,0] as the control group.

Results. Figure 1 shows boxplots of speech duration under DAF for all condition combinations, with the three large groups representing different amount of delay. Within each DELAY group, all combinations of the PIT and FMT are represented as [PIT, FMT]. The RM-ANOVA results show that the significant main effects were found in DELAY ($F(1, 11) = 12.70, p < .01$) and FMT ($F(1.44, 15.8) = 8.59, p < .01$). There was also a significant interaction between DELAY and FMT ($F(4, 44) = 3.64, p < .001$). The significant simple main effects of FMT on DELAY were found for DELAY at 0 ms ($F(1.69, 18.6) = 11.34, p < .001$) and DELAY at 200 ms ($F(2, 22) = 14.84, p < .001$). A marginally significant simple main effect was also found for DELAY at 100 ms ($F(2, 21.9) = 3.12, p = .06$). The multiple comparisons for DELAY $\times$ FMT showed that significant differences were found for [FMT:Up] > [FMT:0] ($p < .05$) and [FMT:Down] > [FMT:0] ($p < .05$) for DELAY at 0 ms. In contrast, significant differences were found for [FMT:0] > [FMT:Up] ($p < .05$) and [FMT:0] > [FMT:Down] ($p < .05$) for DELAY at 200 ms. Therefore, the effect of FMT were reversed on the speech duration depending on the DELAY. This result indicates that the effect of FMT on speech duration is reversed for defferent DELAYs. The results of the Dunnett test with [0,0] were shown within Fig. 1.

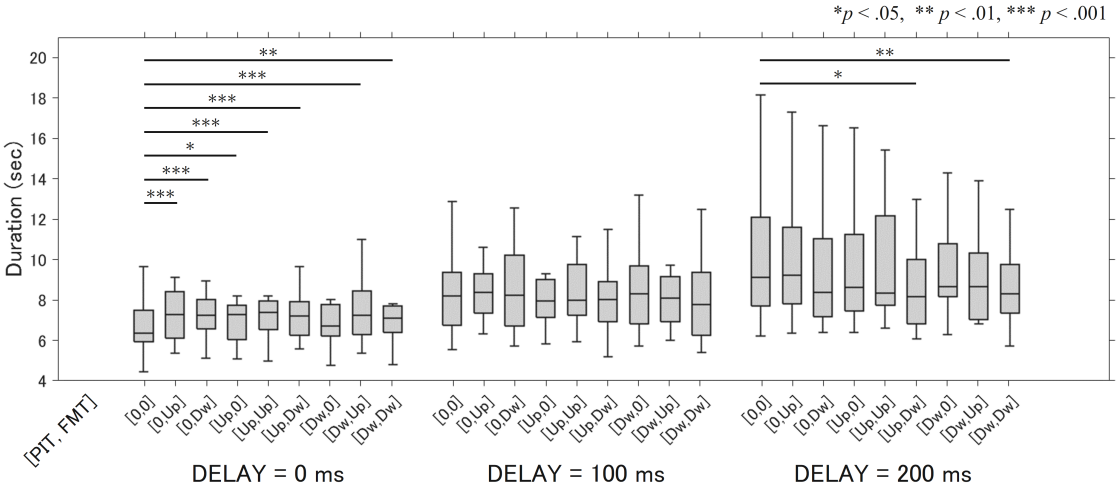


Figure 1: *Speech duration under DAF for all condition combinations. 3 large groups represent different DELAY conditions. Within each group, all combinations of [PIT, FMT] conditions are represented.*

Discussion.

The results of this study showed that speech duration consistently increased as the amount of delay given to the feedback speech under DAF increased. It was also found that the spectral manipulation applied to the feedback speech altered the DAF effect on speech duration. Furthermore, there was a significant interaction between the amount of delay and the spectral manipulation given to the feedback speech. Interestingly, the opposite trend was observed in the condition in which the speaker's own voice was used as feedback: speech duration was shorter when the delay was 0 ms and longer when the delay was 200 ms. This suggests that when auditory feedback is given as the speaker's own voice, the speech duration is more affected by DAF. The main factors that cause changes in speech duration in DAF may include repetition, misarticulation, and increased pauses in speech. These increases in non-fluency in speech are thought to be caused by various errors in the speech production process, such as disturbances in the source production mechanism and articulatory motor control. Therefore, the results suggest that the acoustic individuality of the auditory feedback speech plays an important role in the speech production. Further research based on the results of this study is needed to clarify these relationships.

Acknowledgments. This study was supported by Kawai Foundation for Sound Technology and Music.

References.

Cai, S., M. Boucek, S. S. Ghosh, F. H. Guenther, and J. S. Perkell (2008). "A System for Online Dynamic Perturbation of Formant Trajectories and Results from Perturbations of the Mandarin Triphthong /iau/". In: *Proceedings of the 8th ISSP*, pp. 65–68.

Lee, B. S. (1950). "Effects of Delayed Speech Feedback". In: *The Journal of the Acoustical Society of America* 22.6, pp. 824–826.

Toyomura, A. and T. Omori (2005). "Auditory Feedback Control during a Sentence-Reading Task: Effect of Other's Voice". In: *Acoustical Science and Technology* 26.4, pp. 358–361.

Perception and production are related in novice learners of Mandarin lexical tone

Jennifer Yang[1], Xi Chen[1], Joyce Chung[2], Charles B. Chang[2], Tara McAllister[1]

[1]New York University [2]Boston University

jy3286@nyu.edu, x.chen@nyu.edu, joyce@bu.edu, cc@bu.edu, tkm214@nyu.edu

Introduction. The nature of perception–production relations in second language (L2) learning has been widely studied. Flege's (1995) Speech Learning Model suggests that the accuracy of L2 segmental perception acts as a limiting factor for accuracy in production after a certain proficiency level is reached, and studies such as Bradlow *et al*. (1997) have shown that training in perception can improve production in L2. However, some studies have found no link between L2 perception and production of segmental targets (e.g., Kartushina & Frauenfelder 2014), while others report weak to moderate correlations with large variation between individuals (Nagle 2018).

The present study investigates L2 perception–production relations in lexical tone contrasts, which may behave differently than segmental targets. In languages with lexical tone, differences in the rate of vocal fold vibration (f0) can signal a difference in meaning between words. For example, in Mandarin, four distinctive tones are recognized (high-level tone as T1, high-rising tone as T2, low-dipping tone as T3, high-falling tone as T4). Yang (2012) found a strong correlation between tone production and perception in L2 learners of Mandarin and also reported that production performance was better than perception. As a possible explanation, Yang suggested that accurate production could be achieved through imitation at a relatively superficial phonetic level, whereas accurate perception requires forming a phonological category. However, other studies have reported conflicting findings; for instance, Kirby & Lu Giang (2021) found no well-defined relationship between tone production and perception in L2 learners of Vietnamese.

To shed light on the perception–production relationship in lexical tone learning, the present study acoustically measured English speakers' productions of L2 Mandarin tones and examined the association between acoustically measured production accuracy and performance on a tone contour perception task. Both imitative and non-imitative productions were examined with the goal of understanding both phonetic and phonological types of learning. While most studies to date have focused on speakers with considerable experience learning a tonal L2, the present study fills a gap in the previous literature by focusing on speakers with no previous exposure to tone languages.

Methods. Participants were 65 female native English speakers between the ages of 18-30 who self-reported no previous experience learning a tone language and no history of hearing loss or diagnosis of a disorder affecting speech or language function. Only female participants were included to keep pitch fairly homogeneous across speakers.

Participants received an introduction to Mandarin tones and how they are presented in the alphabetic system *Pinyin* with tone diacritics, and then completed perception and production tasks. The perception task was the pitch-contour perception test (PCPT) from Wong & Perrachione (2007). In the PCPT, participants identified synthesized stimuli as having level, rising, or falling pitch by matching them with arrows representing the pitch contour. Participants then completed two production conditions: an imitative production probe followed by a non-imitative probe. In the imitative probe, participants saw a syllable in *Pinyin* and also heard one of six female native speakers of Mandarin producing the syllable. In the non-imitative probe, participants saw *Pinyin* syllables with no auditory model. Both probes elicited the syllables /ma/, /nai/, /mai/, and /na/ with all four Mandarin tones, twice each in random order. Participants returned to the lab for a second session in which pitch training was provided; the outcome of pitch training was measured as part of a larger study and will not be reported here. However, the imitative and non-imitative probes were repeated at the start of the session, prior to training. Because patterns of performance were similar across the two administrations of the pre-training probes, they have been pooled for the purpose of the analyses reported here.

Acoustic measurements were obtained from each recorded syllable using Praat software. The onset and offset of the vowel in each syllable were marked, and a script (Chang & Yao, 2016) was used to extract f0 at 10 evenly spaced time points along the duration of each vowel. All the f0 measurements were then normalized to a 5-point scale (*T*-value) that characterizes frequencies relative to the highest and lowest f0 measurements (f0max and f0min) for a given participant. Following Chang & Yao (2016), these measurements were used to calculate the mean *T*-value and range of *T*-values in a vowel (MeanT, RangeT), as well as the turning points of contour tones (T2, T3). Finally, the acoustic measures were converted to differences from mean values derived from a sample of 10 female native Mandarin speakers (of whom six were the model talkers used in the training task). Larger difference scores indicate lower similarity with native speakers.

Results. The association between PCPT score and difference from the native-speaker mean can be seen in **Figure 1** for f0 mean and f0 range (left-hand panel) as well as duration and turning point timing (right-hand panel; T2 and T3 only for turning point). Due to the large number of comparisons (28, spanning four tones, four measures, and two production conditions), all *p*-values were corrected for multiple comparisons using the Benjamini-Hochsberg correction as

implemented with the p.adjust() function in R. After this correction, significant correlations with PCPT score were observed for f0 mean difference for T3 in the imitative condition (Pearson's r = -.23) and f0 range difference for T2 and T4 in the non-imitative condition (r = -.34, r = -.38). In addition, a significant correlation between PCPT score and duration difference was observed for T1 in the imitative condition (r = -.23), and significant correlations between PCPT score and turning point timing difference were observed for T2 in both the imitative and non-imitative conditions (r = -.23, r = -.38), as well as for T3 in the non-imitative condition only (r = -.25).

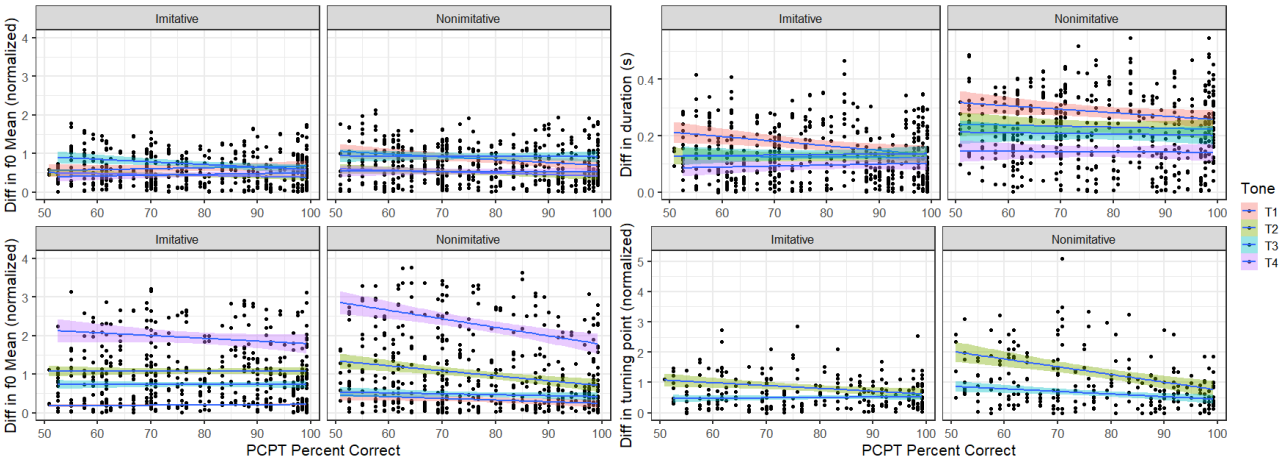


Figure 1: *Associations between PCPT score and difference from native speaker mean in four acoustic measures (f0 mean, f0 range, duration, turning point timing) for tones T1-T4 in imitative and non-imitative conditions.*

Discussion. The results of this study present several interesting points of comparison relative to previous research. First, our findings support the hypothesis that perceptual acuity for pitch affects the ability to produce lexical tone contours in the earliest stages of learning. While the observed perception–production correlations were generally small (the highest observed correlation had a Pearson's r of -.38), they were uniform in having a negative direction, indicating that individuals who scored higher on the PCPT produced tones that were acoustically more native-like. Second, while significant correlations were observed for both the imitative and non-imitative conditions, correlations were generally stronger in the non-imitative condition. Previous research has suggested that imitative production can be accomplished using a superficial phonetic level of processing, whereas non-imitative production requires a phonological level of encoding. The present findings suggest that pitch perception ability could play a more prominent role in generating stored phonological representations of tones (which are then drawn upon for non-imitative production) than in imitating tones at a phonetic level, which is broadly consistent with Yang (2012). Finally, while the primary goal of this study was to examine perception–production relations, it is also of interest to examine differences in the relative accuracy of production across tones and acoustic measures. It was noted that f0 range was associated with an especially large magnitude of error, particularly for T4. Previous research has highlighted T3 as the most challenging tone to produce, and in the present study it was associated with a large magnitude of error relative to other tones for f0 mean. However, T3 was associated with a smaller magnitude of error than T2 and T4 with respect to f0 range, as well as a smaller magnitude of error than T1 with respect to duration and T2 with respect to the timing of the turning point. Future work will examine the relationship between the acoustic measures of tone production obtained here and native listeners' perceptual ratings of accuracy collected as part of the larger study.

References

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, 101(4), 2299-2310.

Chang, C. B., & Yao, Y. (2016). Toward an understanding of heritage prosody. *Heritage Language Journal*, 13(2), 134-160.

Flege, J. E. (1995). Second language speech learning. In W. Strange (Ed.), Speech perception and linguistic experience. York Press. 233-272.

Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, 5, 1246.

Kirby, J., & Lư Giang, Đ. (2021). Relating production and perception of L2 tone. In R. Wayland (Ed.), Second language speech learning: Theoretical and empirical progress. Cambridge University Press. 249-272.

Nagle, C. L. (2018). Examining the temporal structure of the perception–production link in second language acquisition: A longitudinal study. *Language Learning*, 68(1), 234-270.

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, 28(4), 565-585.

Yang, B. (2012). The gap between the perception and production of tones by American learners of Mandarin – An intralingual perspective. *Chinese as a Second Language Research*, 1(1), 33-53.

Auditory Vowel Discrimination in Middle Childhood Compared to Adulthood

Katharina Polsterer[1], Nikki Hoekzema[1], Xingfeng Yang[1], Thomas Tienkamp[1], Teja Rebernik[1], Hedwig Sekeres[1], Reihaneh Amooie[1], Raoul Buurke[1], Wietse de Vries[1], Liyang Wang[2], Frank Tsiwah[1], Martijn Wieling[1], Defne Abur[1]

[1]University of Groningen, [2]University of California, Berkeley
k.m.polsterer@rug.nl

Introduction. Auditory discrimination is considered a crucial factor in speech development during childhood (Guenther 2016). On the one hand, it plays a major role in speech perception and comprehension of the environment (Sussman 2001). On the other hand, self-produced speech sounds need to be discriminated through acoustic features (both on the articulatory and vocal levels) when monitoring and fine-tuning speech production (Guenther 2016). The Directions into Velocities of Articulators model (DIVA; Guenther 2016) proposes that auditory targets for specific speech sounds are formed and become refined during speech development. Accordingly, auditory discrimination is hypothesized to improve with age, suggesting that adults can detect smaller differences in acoustic features than children. In line with this hypothesis, auditory discrimination of pure tone frequencies has been reported to generally improve and become less variable between speakers from age 6 to 11 (Moore et al. 2008). Further, there is evidence of greater variability in auditory discrimination of self-produced pitch (vocal-level feature) in children aged 6 to 11 compared to adults, with children being either less or similarly sensitive to pitch differences (Heller Murray & Stepp 2020). However, prior work proposes separate mechanisms for vocal-level and articulatory-level aspects of DIVA (Kapsner-Smith et al. 2023; Lester-Smith et al. 2020). The objective of the current study was to characterize auditory discrimination of specifically *vowels* (articulatory-level feature) in middle childhood (aged 6–11 years) and adulthood in self-produced speech.

Methods. A total of 94 native Dutch speakers participated in this study after providing written consent in accordance with the University's Research Ethics Review Board. For underage participants (aged 6–11 years), parents or legal guardians provided written consent for the study. The study took place in a sound-attenuated booth situated inside a mobile research van (Wieling et al. 2023). All participants passed a pure-tone audiometric screening, evaluating whether they were able to hear frequencies of 250 to 8,000 Hz at 25 dB hearing level binaurally. Participants were recruited across two age ranges: children in middle childhood (aged 6–11 years; n=49; 25 girls, 24 boys) and adults (aged 20–38 years; n=45; 24 women, 20 men, 1 non-binary). Participants wore an over-the-ear microphone (Shure MX153; positioned 7 cm away from the mouth) and headphones (Sennheiser HD 280 Pro). Auditory discrimination was measured using a perceptual just-noticeable-difference (JND) task (e.g., Lester-Smith et al. 2020; Villacorta et al. 2007). Through this task, we assessed the participants' perception of changes in their vowel productions. The experiment comprised two steps: (1) speech production, followed by (2) a perceptual JND task using the recorded production. Specifically, the participants produced three repetitions of three Dutch words in a random order, namely <daar> /daːr/ 'there', <deur> /døːr/ 'door', and <duur> /dyːr/ 'expensive', when they were prompted on the screen in front of them. The participants were instructed to prolong the vowels for approximately 1 second. Audapter (v2.1.012; Cai et al. 2008) was used to record all productions and select the <deur> production with the median F_1 (measured at the mid-point of the vowel) for the next step. The other two words were elicited so that the participants did not know that <deur> was the word used for the JND task. The participants subsequently listened to trials with pairs of the selected recording presented at 75 dB SPL, which were separated by a 500-ms inter-stimulus interval. The participants were asked to indicate whether the vowels in the pair sounded the same or different. In 20% of the trials, listeners heard the original <deur> production twice ('catch' trials). For the rest of the trials, listeners heard the original <deur> production and a version of the production with F_1 modification. Initially, F_1 was increased by 40% relative to the original production. For example, if the median <deur> production had an F_1 of 440 Hz, a participant would first hear this median production compared to the same production with an F_1 of 616 Hz (+40%). Subsequently, the F_1 of the second stimulus increased or decreased by 3% in the pair after that, depending on the listener's response. This allowed us to determine the auditory discrimination threshold (ADT) in percentage-changed at which participants were still able to hear the difference between the vowels. This procedure was repeated until 6 reversals (i.e., changes in the direction of the F_1 manipulation) or 60 trials were completed.

Results. The mean auditory discrimination threshold (ADT) was 49% (sd=26) among children, whereas it was 33% (sd=21) in adults. As ADT constitutes the threshold of F_1 deviation at which the difference to the reference could still be perceived, a lower value reflects more precise auditory discrimination. A linear regression model revealed that the group

difference in ADT was significant (β=-0.433, p=0.001), indicating lower ADT in children than adults as presented in **Figure 1**. Additionally, Levene's Test did not indicate group differences in variance of ADT (F=1.715, p= 0.194).

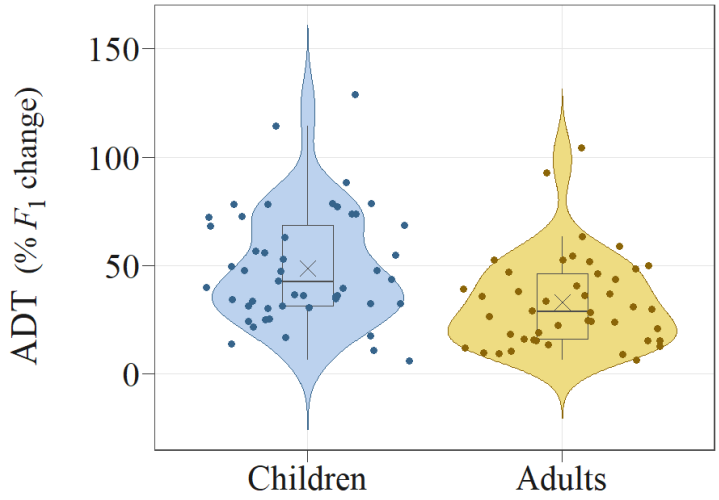


Figure 1: *Auditory discrimination thresholds in % change from the original median <deur> F_1, with means per group (x) and individual data points.*

Discussion. In this study, we examined auditory discrimination of F_1 in self-produced vowels in middle childhood (6–11 years of age) and adulthood (20–38 years of age). The results showed a significant difference in auditory discrimination between age groups, with children overall discriminating vowels less precisely than adults. This lends support to the DIVA model (Guenther 2016) suggesting that in middle childhood, auditory target regions for vowels are still developing. The current finding corresponds to the patterns previously observed for pitch, which showed that, overall, children were either less or similarly sensitive to pitch differences compared to adults (Heller Murray & Stepp 2020). Thus, articulatory-level and vocal-level auditory discrimination seem to be developing in parallel during middle childhood, while the underlying mechanisms might differ (Kapsner-Smith et al. 2023; Lester-Smith et al. 2020). Furthermore, in the current study, variability in auditory discrimination of vowels did not differ between the groups of children and adults (as measured through Levene's Test). This suggests that articulatory-level discrimination remains variable between speakers, even in adulthood. This result differs from previous findings in the pitch domain: children in middle childhood showed greater between-speaker variability in auditory discrimination of pitch than adults (Heller Murray & Stepp 2020). As both auditory discrimination of vowels and vowel production reflect auditory target representations, according to DIVA (Guenther 2016), less precise auditory discrimination might be related to more production variability. Therefore, one possible next step is to investigate the link between auditory discrimination of vowels and vowel production variability in speech development.

References

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2008). *A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/* [Paper presentation]. 8th International Seminar on Speech Production, Strasbourg, France.

Guenther, F. H. (2016). *Neural control of speech*. MIT Press.

Heller Murray, E. S., & Stepp, C. E. (2020). Relationships between vocal pitch perception and production: A developmental perspective. *Scientific Reports*, *10*(1), 3912.

Kapsner-Smith, M. R., Abur, D., Eadie, T. L., & Stepp, C. E. (2023). Test-retest reliability of behavioral assays of feedforward and feedback auditory-motor control of voice and articulation. *Journal of Speech, Language, and Hearing Research*, 1–15.

Lester-Smith, R. A., Daliri, A., Enos, N., Abur, D., Lupiani, A. A., Letcher, S., & Stepp, C. E. (2020). The relation of articulatory and vocal auditory-motor control in typical speakers. *Journal of Speech, Language, and Hearing Research*, *63*(11), 3628–3642.

Moore, D. R., Ferguson, M. A., Halliday, L. F., & Riley, A. (2008). Frequency discrimination in children: Perception, learning and attention. *Hearing Research*, *238*(1-2), 147–154.

Sussman, J. E. (2001). Vowel perception by adults and children with normal language and specific language impairment: Based on steady states or transitions? *Journal of the Acoustical Society of America*, *109*(3), 1173–1180.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America*, *122*(4), 2306–2319.

Wieling, M., Rebernik, T., & Jacobi, J. (2023). *SPRAAKLAB: A mobile laboratory for collecting speech production* [Paper presentation]. 20th International Congress of Phonetic Sciences, Prague, Czech Republic.

Speaking-induced Middle Ear Muscle Reflex (MEMR): suppression of auditory feedback during self-vocalization

Hayo Terband[1], Caroline Cross[1,2], Joel Berger[3], Shawn Goodman[1]

[1]Department of Communication Sciences and Disorders, University of Iowa, Iowa City IA, USA
[2]Hearing Associates, Mason City IA, USA
[3]Department of Neurosurgery, University of Iowa, Iowa City IA, USA

hayo-terband@uiowa.edu, ccross@hearingassociatesmc.com, joel-berger@uiowa.edu, shawn-goodman@uiowa.edu

Introduction.
Corollary discharge (CD) is an umbrella term for brain functions that allow animals to differentiate external from self-generated sensory signals and encompasses both lower- and higher-order mechanisms, depending on their function (Crapse & Sommer 2008; Sperry 1950; Holst & Mittelstaedt 1950). Lower-order mechanisms concern the control of sensation by the Central Nervous System (CNS) and include sensory filtration and reflex inhibition; higher-order mechanisms concern the control of action and perception and include sensory analysis/stability and sensorimotor learning/planning (Crapse & Sommer 2008).
One higher order mechanism relevant to human speech production is speaking-induced suppression (SIS), the phenomenon of a reduced response in auditory cortex to self-produced compared to externally-produced speech (Numminen & Curio 1999; Houde et al. 2002; Greenlee et al. 2011). SIS is thought to be triggered by the efference copy from motor cortex containing a forward prediction of the sensory consequences of the motor program (Knille et al. 2019; Ylinen et al. 2015) and/or the sensory goals associated with the motor plan (Niziolek et al. 2013). The mechanism is thought to play an important role in error detection and –correction, and speech-motor learning.
Largely ignored in the field of speech production but well-studied in audiology are the lower-order CD mechanisms, which involve the two major efferent feedback pathways to the auditory periphery: the middle ear muscle reflex (MEMR) reflex and the medial olivocochlear reflex (MOCR). In particular MEMR is of interest in the context of SIS. MEMR involves the contraction of the intratympanic muscles, which increases the stiffness of the ossicular chain, thereby altering the acoustic impedance (Metz 1952), particularly below about 1.5kHz. This reduces the input to the cochlea at these frequencies and, in quiet environments such as experimental lab conditions, subsequently the response in the auditory cortex (Herrmann et al. 2020). Clinical MEMR thresholds to external stimuli are relatively high; about 75dB-SPL for noise and 90dB-SPL for pure-tones in healthy listeners (Liberman & Guinan 1998). Perhaps because of this, a common misconception is that MEMR is irrelevant for normal conversational speech, with voice levels of 60-70dB-SPL. However, EMG data has shown that MEMR can also occur without acoustic stimulation during (and in anticipation of) vocalization at normal vocal effort (Borg & Zakrisson 1975). Furthermore, the effect is stronger during self-vocalization than when presented with external speech (Borg & Zakrisson 1975). SIS is determined by subtracting the magnitude of cortical responses during speaking from the response magnitude during listening to playback of the same speech signal. Since it alters the signal input from the periphery, MEMR thus forms a major confound. The current study features a novel, non-invasive method to measure MEMR that allows isolation and quantification of the MEMR component of SIS.

Methods. 15 young adult speakers of American English (11 females, 4 males; age range = 18–25 years) with normal hearing and speech participated in the study. The first five participated in pre-pilot and development. Participants were seated in a sound-treated booth; stimuli were presented, and ear canal pressure measured binaurally using a 2-channel probe-microphone system (Etymotic ER10X). The experiment consisted of 110 trials, each consisting of the conditions *Listen*, in which subjects listened to the recording of their own voice playing back the word "cup" five times, and *Speak*, in which subjects were visually cued to produce the word "cup" five times, with a 2.5s inter-stimulus interval. During both *Listen* and *Speak* trials, a train of low-level clicks were played continuously. Sound pressure levels of stimuli were equivalent. The time-course and magnitude of MEMR responses were quantified by measuring the changing amplitudes of the click sounds reflected by the eardrum during the inter-stimulus intervals and during two baseline conditions preceding the *Listen* and *Speak* conditions in which only the low-level click train were played continuously (**Figure 1**).

Results. Figure 2 presents changes in recorded waveform magnitude averaged over five 100ms time windows with fitted MEMR activation curves during the inter-stimulus intervals in *Listen* and *Speak* conditions compared to baselines.

Discussion. The results show that MEMR can be assessed non-invasively by examining the sound pressure of acoustic clicks in the ear canal recorded in-between speech stimuli. Results were consistent with MEMR being activated prior to the onset of speech, but not prior to the playback of the recorded utterances. Changes in pressure were also stronger for self-generated sound. These findings have implications for research into SIS. As MEMR reduces the excitation amplitude

in the cochlea and subsequently the response in auditory cortex, SIS needs to be corrected for it. Experiments involving simultaneous measurements of MEMR and cortical activation for self-vocalizations are being prepared.

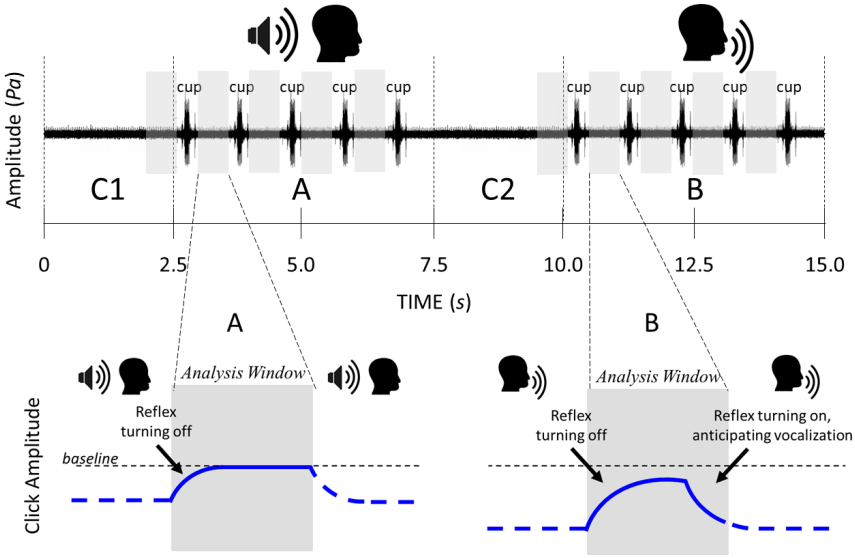


Figure 1: *Experimental paradigm and predicted results. In response to external sound (A), the reflexes will turn on after onset, with a delay of ~100ms. The reflexes will turn off after the sound stops with the same delay. In response to self-produced sound (B), the reflexes will be activated more strongly and prior to onset.*

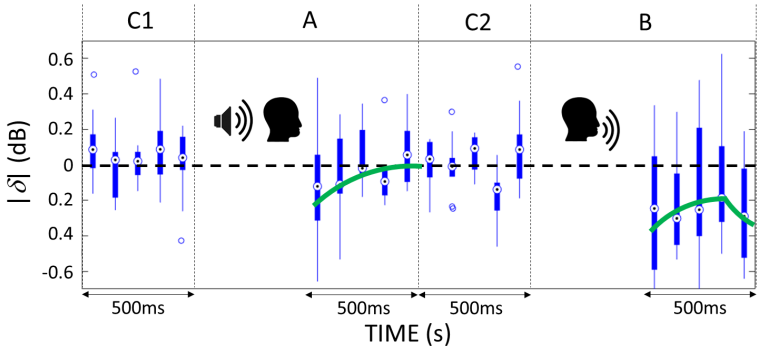


Figure 2: *Mean (n=10) changes in recorded waveform magnitude (|δ|) with fitted MEMR activation curves (in green) over the 500ms analysis windows in the conditions* Listen *(A),* Speak *(B) compared to baselines (C1 & C2).*

References

Borg, E., & Zakrisson, J. E. (1975). "The activity of the stapedius muscle in man during vocalization". In: *Acta oto-laryngologica*, 79.3-6, pp. 325-333.

Crapse, T. B., & Sommer, M. A. (2008). "Corollary discharge across the animal kingdom". In: *Nature Reviews Neuroscience*, 9.8, pp. 587-600.

Greenlee, J. D. W., Jackson, A.W., Chen, F., Larson, C.R., Oya, H., Kawasaki, H., ... & Howard III, M. A. (2011). "Human Auditory Cortical Activation during Self-Vocalization". In: *PLoS ONE*, 6.3, pp. e14744.

Herrmann, B., Augereau, T., & Johnsrude, I. S. (2020). "Neural responses and perceptual sensitivity to sound depend on sound-level statistics". In: *Scientific Reports*, 10.1, pp. 9571.

Holst, E. von. & Mittelstaedt, H. (1950). "The reafference principle". In: *Naturwissenschaften*, 37, pp. 464–467.

Numminen, J., & Curio, G. (1999). "Differential effects of overt, covert and replayed speech on vowel-evoked responses of the human auditory cortex". In: *Neuroscience letters*, 272.1, pp. 29-32.

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). "Modulation of the auditory cortex during speech: an MEG study". In: *Journal of cognitive neuroscience*, 14.8, pp. 1125-1138.

Knolle, F., Schwartze, M., Schröger, E., & Kotz, S. A. (2019). "Auditory predictions and prediction errors in response to self-initiated vowels". In: *Frontiers in Neuroscience*, 13, pp. 1146.

Liberman, M. C., & Guinan Jr, J. J. (1998). "Feedback control of the auditory periphery: anti-masking effects of middle ear muscles vs. olivocochlear efferents". In: *Journal of communication disorders*, 31.6, pp. 471-483.

Metz, O. (1952). "Threshold of reflex contractions of muscles of middle ear and recruitment of loudness". In: *AMA Archives of Otolaryngology*, 55.5, pp. 536-543.

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). "What does motor efference copy represent? Evidence from speech production". In: *Journal of Neuroscience*, 33.41, pp. 16110-16116.

Sperry, R. (1950). "Neural basis of the spontaneous optokinetic response produced by visual inversion". In: *Journal of Comparative and Physiological Psychology*, 43, pp. 482–489.

Ylinen, S., Nora, A., Leminen, A., Hakala, T., Huotilainen, M., Shtyrov, Y., ... & Service, E. (2015). "Two distinct auditory-motor circuits for monitoring speech production as revealed by content-specific suppression of auditory cortex". In: *Cerebral Cortex*, 25.6, pp. 1576-1586.

Analysing the vocal tract front-back relationships

Antoine Serrurier

Clinic for Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital and Medical Faculty of the RWTH Aachen University, Aachen, Germany

aserrurier@ukaachen.de

Introduction. During speech production, the buccal and pharyngeal parts of the vocal tract vary significantly. For physiological reasons, they present however some degrees of correlation. As emphasised by Stevens (2000), "*a narrowing in one part of the vocal tract will automatically be accompanied by a widening in other parts, since there is a constraint that the vocal tract volume is roughly constant*" (p. 261). In a recent study, a deep neural network trained to classify vowels from the shape of the vocal tract ended up focusing almost exclusively on the buccal part of the vocal tract to take a classification decision, even for vowels with a constriction, crucial for the acoustics, located in the pharyngeal part (Serrurier, 2023). The current study aims at better understanding the front-back relationships by trying to predict the shape of the back part of the vocal tract from the shape of the front part. It relies on an articulatory modelling approach of the vocal tract for which control parameters are estimated solely from the front part.

Methods. Static midsagittal Magnetic resonance imaging (MRI) data from speakers sustaining phonemes representative of the articulatory repertoire of their native language have been considered. Altogether, the data consist of 41 adult speakers (18 females, 23 males) from 3 different native languages, representing altogether 1948 midsagittal images of the vocal tract. The number of images per speaker varies between 27 and 143. For each image, the contours of the articulators surrounding the vocal tract from the glottis to the lips have been manually delineated and re-interpolated with a fixed number of points between cross-speaker anatomical landmarks, with 1036 points altogether. The contour coordinates corresponding to an image are referred to as an articulation. Two illustrations are visible in Figure 1. Further details on these data and the contours can be found in Serrurier & Neuschaefer-Rube (2023).
For each speaker, a model of the shape of the vocal tract by means of Principal Component Analysis (PCA) has been developed. With the model, each articulation can be approximated by a set of fixed eigenvectors, corresponding to the components, weighted by articulation-dependant control parameters. Any component explaining more than 1% of the data variance has been retained, leading to models made of about 10 components and usually explaining more than 95% of the data variance. Any new articulation can be approximated by the model by obtaining the control parameters by linear regression of the contours on the eigenvectors and then combining the calculated control parameters with the eigenvectors. Further, the control parameters can be estimated from a restriction of the contours and the eigenvectors, typically limited to points considered as belonging to the front part of the vocal tract. By combining the calculated control parameters with the full eigenvectors, the full articulation can be estimated. The smaller the portion of what is considered as front part of the vocal tract used for control parameter estimation, the larger the error on the control parameter and on the recovered articulation.
This approach has been used on a leave-one-articulation-out scheme: for each speaker, a test articulation is left out and the model built on the other articulations. The left articulation is used for test purposes and the process is repeated until all articulations are used once as test articulation; the overall results are reported. For testing, the portion considered as the front part of the vocal tract, used to estimate the control parameters, has been continuously varied from the position A to the position B in Figure 1 (front part in green, back part in red). The reconstruction error has been calculated in terms of Root-Mean-Squared (RMS) error on the remaining part of the vocal tract, considered as the back. A baseline error has been calculated on the back part of the vocal tract obtained with the control parameters calculated from the full articulation contours; it represents the prediction capacity of the model. The RMS baseline has been subtracted from the RMS error, leading to a ΔRMS error, so that the optimal estimation of the control parameters leads to a ΔRMS error of 0.

Results. The results per speaker are visible on Figure 1. As expected, using the lips and the jaw to predict the shape of the rest of the vocal tract leads to very large errors. Logically, the error decreases with the increase of the portion considered as the front part of the vocal tract. Preliminary analyses show that most speakers curves follow a decrease in one or two steps, as illustrated by the yellow and red curves on Figure 1. When there is one step, this region tends to correspond to the tongue dorsum. When there are two steps, the regions tend to correspond to the tongue tip/blade and the tongue dorsum, with variability between speakers. In other words, when the tongue dorsum and/or the tongue tip/blade are estimated from the front part of the vocal tract, which does not include these parts, it tends to generate large reconstruction errors.

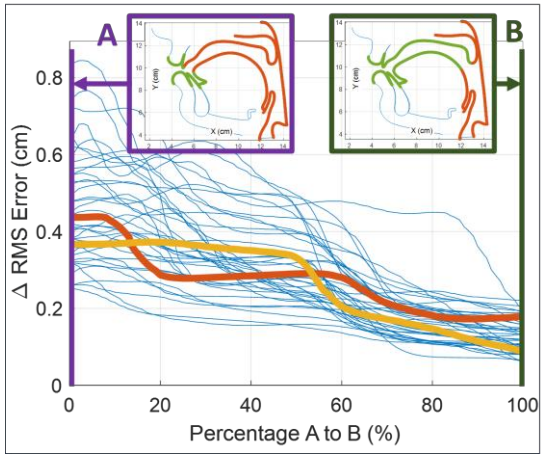


Figure 1: ΔRMS reconstruction error of the back part of the vocal tract (in red on articulations A and B) vs. percentage of the vocal tract considered as front part (green) for 40 out of 41 speakers (one outlier speaker not displayed); the curves for two typical speakers are displayed in thick yellow and red.

Discussion. The results suggest that estimating the back part of the vocal tract including the tongue dorsum and/or the tongue tip/blade from the front part of the vocal tract, not including these sections, leads to significantly larger errors than including them in the front part of the vocal tract. As a consequence, it suggests that the tongue tip/blade and the tongue dorsum tend to be more uncorrelated with the back part of the vocal tract than other sections of the tongue. This seems coherent with the fact that the tongue tip, and to a lower extent the tongue dorsum, can move independently from the back part of the vocal tract, for example to achieve an alveolar or a velar stop. On the contrary, a large frontward-backward movement of the tongue body may be more correlated with variations in the pharyngeal region.

One technical limitation in the current approach is that the decreasing error observed in the curves while progressing from A to B positions can be ascribed to two sources: (1) the better estimation of the control parameters due to a larger part of what is considered as the front part of the vocal tract and (2) the lower reconstruction error calculated on a smaller part of the vocal tract considered as the back part. Further complementary analyses may consider a fixed part of what is considered as the back part of the vocal tract to calculate the error. An extension of the extreme positions A and B can also be considered in the future.

Further analyses are also necessary to deal with the large inter-speaker variability observed in the results. While some common trends have been found, speakers still show different behaviours. In addition, the results still need to be interpreted considering the biomechanics of the vocal tract, as the front-back correlations are linked with physiological constraints. Complementarily, a finer analysis of the reconstruction error per sub-regions can provide significant insights: one may indeed hypothesise that the pharyngeal area, partly dependant on the tongue shape and position, may be more correlated with the front of the vocal tract than for instance the larynx height.

Despite the current limitations, the exploratory method presents an original approach to evaluate the front-back relationships in the vocal tract. This may help to understand motor planning, as it characterises to which extent the shape of the back part of the vocal tract can be derived from the front part. Moreover, assuming that speakers are more aware of movements in the front part of the vocal tract, it can provide in the future new insights on the (re-)education of full articulations driven by strategies focusing on the front part of the vocal tract. In more practical terms, the study provides a quantitative evaluation of the level of information available for the back part of the vocal tract captured by articulatory tracking systems focusing on the front part of the vocal tract such as Electromagnetic Articulography (EMA). It can also provide guidelines where to locate EMA coils to capture the maximal articulatory information for the whole vocal tract. More generally, it quantifies how far measures from the front part of the vocal tract can be enough to characterise the full vocal tract.

Acknowledgements. Parts of this research project were supported by the START-Program of the Faculty of Medicine, RWTH Aachen University.

References

Serrurier, A. (2023). Can Deep Learning help to understand speech production mechanisms? *Proc. ESSV 2023*, 181-188.

Serrurier, A., & Neuschaefer-Rube, C. (2019). Morphological and acoustic modeling of the vocal tract. The Journal of the Acoustical Society of America, 153(3), 1867-1886.

Stevens, K. (2000). Acoustic Phonetics. MIT Press.

Acoustic cues to lexical stress in Bulgarian

Milena Milenova

Sofia University

mamilenova@uni-sofia.bg

Introduction. The lexical stress in Contemporary Standard Bulgarian (CSB) is free and differentiates minimal pairs differing in lexical or grammatical meaning. The acoustic correlates of CSB lexical stress have not been investigated systematically. Studies are scarce, if any, and neither methodology, nor measured values regarding the existing claims are reported. According to Boyadzhiev & Tilkov (1999) CSB stress is manifested mainly through intensity and is therefore defined as dynamic. Duration and fundamental frequency (F0) are also considered important for stress marking. None of these acoustic dimensions is considered the solely responsible cue to word stress. Rather, stress is understood as the result of the joint effect of multiple acoustic properties whose effect may vary as their role as stress markers may interplay with syllable position within the word, as well as word position within the utterance. Vowel quality may also play some role, however its impact may be limited to the vowels /a/ and /ɔ/ which when unstressed undergo reduction and are realised respectively as [ɜ] and [u] (Andreeva *et al*. 2013; Dokovova *et al*. 2019; Sabev, 2023). This paper is part of an on-going research project whose aim is to re-examine the existing views on the acoustic and auditory aspects of stress in CBS. The present pilot study explores the relative importance of intensity, F0 and vowel duration as stress correlates in relation to syllable position. In addition, it investigates the effect of vowel quality for stress marking by comparing the stress-induced changes in F1 and F2 of the vowels /a/ and /ɔ/ which undergo categorical reduction when unstressed and of the vowels /ɜ/ and /u/ which are not subject to phonological reduction. The role of these acoustic parameters as stress exponents in CBS is examined in light of the reported in Gordon & Roettger (2017) crosslinguistic evidence showing that the robustness of these particular acoustic cues is in order of magnitude as follows: duration, F0, intensity, formant structure.

Methods. Six female speakers of CSB participated in a production experiment. Following Arvaniti (2000) the stimuli for the experiment were the disyllables [ˈpVpV] and [pVˈpV] produced with the vowels [a, ɜ, ɔ, u] in the carrier sentence ‘’*Pepi* _____ *ˈpali*’ (‘Pepi puts the _____ on fire’). These disyllables are non-words in CSB but were chosen as they allow for comparisons within and across test words. The speech material consisting of 6 repetitions per speaker was recorded digitally at 44.1 kHz sampling rate and 16-bit resolution using a portable Marantz 660 Flashcard recorder and a RØDE NT55 condenser microphone. For the purpose of the study measurements of mean intensity, mean F0, vowel duration, mid F1 and mid F2 for each vowel were obtained using Praat (Boersma & Weenink, 2023). All measures were collected from adjacent syllables and from syllables with identical position in the word.

Results. The collected data were explored by conducting separate RM ANOVAs for each acoustic parametre with Stress, Position and Vowel as the within subject factors. The results revealed that all stressed vowels were significantly longer and had significantly higher intensity than their unstressed counterparts irrespective of their position in the word. The efficacy of F0 and formant frequencies depended on vowel and syllable position. F0 distinguished [a+str] from [a-str] in initial syllables and [u+str] from [u-str] in both initial and final syllables, but was not efficient for distinguishing stressed and unstressed syllables with [ɔ] and [ɜ]. F1 worked well for the distinction between [ɔ+str] and [ɔ-str], as well as for [a+str] and [a-str] in both syllable positions, less so for [ɜ+str] and [ɜ-str] and not at all for [u+str] and [u-str]. F2 distinguished [a+str] from [a-str] in both initial and final position, and [ɜ+str] from [ɜ-str] in initial position, but did not work for the back rounded vowels.

Discussion. The aim of this pilot study was to explore the relative efficacy of the acoustic measures reported in the literature as stress markers in CSB. The present results are partially in line with the existing claims. The findings corroborate the reported robustness of intensity and duration as stress cues in CSB, but do not support the reliability of F0 as a stress correlate. The low efficiency of F1 and F2 as stress exponents was expected due to the stress-induced phonological vowel reduction in CSB. The present duration and the formant frequencies results are in line with the existing crosslinguistic evidence, however the efficacy of intensity and F0 is reversed here. Moreover, the data for F0 suggest that its efficacy may be vowel dependent.

Considering that the instrumental evidence for both the acoustics and the perception of CSB lexical stress is fairly scarce, and that these results are preliminary, future research will be done including more acoustic parametres and perception tests.

References

Andreeva, B., Barry, W., Koreman, J. (2013). The bulgarian stressed and unstressed vowel system. A corpus study. *Interspeech 2013. Proc. Of 14th Annual Conference of the International Speech Communication Association,*2720–2724. *doi:10.214377.*

Arvaniti, A. (2000). The phonetic of stress in Greek. *Journal of Greek Linguistics* 1, 9-39.

Boersma, P. & Weenink, D. (2023). Praat: doing phonetics by computer, version 6.4.01, http://www.praat.org/.

Boyadzhiev, T., Tilkov, D. (1999). *Phonetica na bylgarskiya ezik*. Veliko Tyrnovo: Abagar.

Dokovova, M., Sabev, M., Scobbie, J., Lickley, R., Cowen, S. (2019). Bulgarian Vowel Reduction in Unstressed Position : an Ultrasound and Acoustic Investigation". *Proceedings of the 19th International Congress of Phonetic Sciences. Australasian Speech Science and Technology Association Inc. Canberra:* 2720–2724.

Gordon, M & Roettger, T. B. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistic vanguard.* 20170007.

Sabev, M. (2023). Unstressed vowel reduction and contrast neutralisation in western and eastern Bulgarian: A current appraisal. *Journal of Phonetics*, 99, 101-242.

Silber-Varod, V., Sagi, H., Amir, N. (2016). The acoustic correlates of lexical stress in Israeli Hebrew. *Journal of phonetics*, 56, 1-14.

An exploration of pitch in Afro-Mexican Spanish

Gilly Marchini[1]

[1] *University of Edinburgh*

G.E.M.Marchini@sms.ed.ac.uk

Introduction. This paper documents the role of pitch in Afro-Mexican Spanish, an under-researched variety of Spanish spoken by isolated communities of African heritage in the South-West of Mexico (Oaxaca and Guerrero). Results suggest that whilst phrase-level pitch is employed as part of intonation, unique peak alignment patterns emerge. It considers how features diverge from non-Afro Spanishes, and their theoretical bearing upon pitch anchoring processes. The research questions ask:

RQ1. What is the variation in prenuclear and nuclear pitch accent realisation in Afro-Mexican Spanish?

RQ2. What is the distribution of tonic versus post-tonically aligned peaks?

RQ3. How do features diverge from prosodic descriptions of non-Afro Mexican Spanishes?

Methods. Data was elicited through sociolinguistic interviews recorded in the field (Costa Chica, Mexico). Interviews were conducted in a group setting with a community liaison present at all times. For this paper, 122 broad focus, declarative, Intonational Phrases (IPs) were analysed from a 51-year-old, female speaker of Afro-Mexican Spanish (from Punta Maldonado, Oaxaca, Mexico). Narrow focus utterances were excluded due to the likelihood of tonic peaks in this condition (Martín Butragueño 2006). Speech was recorded on a ZOOM recorder and a head-mounted microphone three inches from the mouth. Data was segmented using the MAUS aligner and manually corrected. Phrases were annotated according to Sp_ToBI protocol (Mota et al. 2011) with ToBI labels extracted via Praat script. Here within, L+H* denotes tonically aligned peaks, i.e., those reached within the stressed syllable, and L+>H* post-tonic peaks, i.e., those reached in the following syllable. Statistical analysis and plots were run and created in R (R Core Team 2022).

Results. L* and L+H* accounted for the majority of prenuclear pitch accents (28.98% each), followed by H* (19.79%) and L+>H* (13.07%). H+L* & L*+H accounted for 8.13% and 1.03% respectively. Nuclear accents were coded into *circumflex* versus *other*, with circumflex accents accounting for 72% of configurations. The circumflex accent was equally likely in IPs containing more than one prosodic word as those containing one (t = -0.3791, p > .01).
With all peaks pooled together, tonic peaks (L+H*) were more common than post-tonic peaks (L+>H*). However, this varied according to syllable aperture: L+H* was significantly more likely in open syllables than closed (t = 2.8622, p <.001). Comparisons also revealed an interaction between syllable structure and the following nasal: in closed syllables, i.e., with coda /n/, e.g., *descendiente*, tonic peaks accounted for 90.5% of rises. In open syllables, however, i.e., with /n/ as the following onset, e.g., *mexicano*, 100% of peaks were post-tonic (Figure 1).

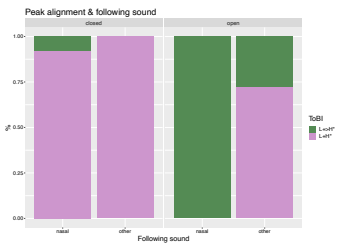


Figure 1: *Prenuclear peak realisation across nasal contexts and syllable aperture.*

Discussion. Results indicate that L* and L+H* account for the majority of prenuclear pitch accents and the circumflex accent in nuclear position. These features thus align with descriptions of non-Afro Mexican Spanishes where such features signal broad focus declarative conditions as part of phrase-level intonation (Mota et al. 2011; Martín Butragueño 2003; Martín Butragueño 2004; Martín Butragueño 2006; Martín Butragueño 2019; O'Rourke 2012). As such, Afro-Mexican Spanish diverges from Afro-Hispanic language features where pitch may signal distinctions in syllable stress (Hualde and Schwegler 2007; Lipski 2004; Lipski 2008).

Nonetheless, whereas post-tonic peaks are common in non-Afro Mexican varieties (Martín Butragueño 2003; Willis 2005), the data here point to tonic peak alignment, albeit to a greater extent in closed syllables. A suprasegmental-segmental interaction also emerges whereby, if present in the segmental string, peaks align on following nasals regardless of intervening syllable boundaries. Whilst tonic peak alignment is common across Afro-Hispanic language, the role of nasal is unattested. This therefore suggests that nasals may act as the pitch anchor in this variety, and that tonal alignment is not always constrained by syllable boundaries, but rather the segmental string.

This raises important theoretical questions surrounding the suitability of the Segmental Anchoring Hypothesis (SAH), in which tonal movements align with syllabic units regardless of their segments. Instead, we consider the following options: firstly, it may be that a lax, dialect-specific SAH emerges due to an underlying phonological feature, e.g., nasality or sonority (Atterer and Ladd 2004). In this way, tonal movements are aligned to syllabic units, as evidenced by the prevalence of tonic peaks across syllable types, yet anchor to nasals when present in the segmental string. Secondly, an articulatory, inter-gestural coordination model may be applicable. It can be theorised that tonal release patterns follow that of the supra-glottal gestures: gestures are tightly coordinated at syllable onset, yet variable and unstable at the syllable offset. As such, pitch offset alignment, here the peak, is variable according to the the phonetics and timings of the coda (Prieto, Van Santen, and Hirschberg 1995; Prieto and Torreira 2007; Prieto 2009). We may theorise that the longer duration of the nasal provides a platform for peak alignment or that, in order to for the rise to be perceptually salient, it must continue throughout the nasal offset, regardless of intervening prosodic boundaries (House 1990).

We are unable to answer the remaining questions at present. Rather, they point to further investigation, and are thus the primary motivation for upcoming control experiments. Nonetheless, they are indicative of the uniqueness of the variety, and the importance of exploring under-researched varieties in order to shine light on theoretical questions surrounding pitch anchoring.

References.

Atterer, Michaela and D. Robert Ladd (2004). "On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German". In: *Journal of Phonetics* 32.2, pp. 177–197. DOI: 10.1016/S0095-4470(03)00039-1.

House, David (1990). *Tonal Perception in Speech*. Lund: Lund University Press.

Hualde, José Ignacio and Armin Schwegler (2007). "Intonation in Palenquero". In: *Journal of Pidgin and Creole Languages* 23.1, pp. 1–31.

Lipski, John M. (2004). "The Spanish of Equatorial Guinea". In: *Arizona Journal of Hispanic Cultural Studies* 8, pp. 115–130.

— (2008). *Afro-Bolivian Spanish*. Frankfurt/Madrid: Vervuert/Iberoamericana.

Martín Butragueño, Pedro (2003). "Hacia una descripción prosódica de los marcadores discursivos: datos del español de México". In: *La tonía: dimensiones fonéticas y fonológicas*. Ed. by Esther Herrera Z. and Pedro Martín Butragueño. Mexico: Colegio de México, pp. 375–402.

— (2004). "Configuraciones circunflejas en la entonación del español mexicano". In: *Revista De Filología Española* 84.2, pp. 347–373.

— (2006). "El estudio de la entonación del español de México". In: *Haciendo lingüística. Homenaje a Paola Bentivoglio*. Ed. by Mercede Sedano, Adriana Bolívar, Martha Shiro, and Antonio Torres, pp. 105–125.

— (2019). "Aproximación a la entonación del español de la ciudad de Oaxaca, México: hacia una geoprosodia". In: *Moenia* 25, pp. 539–596.

Mota, Carme de la, Pedro Martín Butragueño, Pilar Prieto, and Pompeu Fabra (2011). "Mexican Spanish Intonation Mexican Spanish Intonation". In: *Transcription of Intonation of the Spanish Language*. Ed. by Pilar Prieto and Paolo Roseano. Munich: LINCOM Europa, pp. 319–359.

O'Rourke, Erin (2012). "Intonation in Spanish". In: *The Handbook of Hispanic Linguistics*. John Wiley & Sons, Ltd. Chap. 9, pp. 173–191. DOI: https://doi.org/10.1002/9781118228098.ch9. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118228098.ch9.

Prieto, Pilar (June 2009). "Tonal alignment patterns in Catalan nuclear falls". In: *Lingua* 119.6, pp. 865–880. DOI: 10.1016/j.lingua.2007.11.014.

Prieto, Pilar and Francisco Torreira (2007). "The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish". In: *Journal of Phonetics* 35.4, pp. 473–500. DOI: 10.1016/j.wocn.2007.01.001.

Prieto, Pilar, Jan Van Santen, and Julia Hirschberg (1995). *Tonal alignment patterns in Spanish*. Tech. rep., pp. 429–451.

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna, Austria.

Willis, E (2005). "Tonal levels in Puebla Mexico Spanish declaratives and absolute interrogatives". In: *Theoretical and experimental approaches to Romance linguistics*. Ed. by Randall Gess and Ed Rubins, pp. 351–363.

Mandarin Chinese tonal coarticulation in the production of learners with an atonal L1

Kornélia Juhász, Huba Bartos

HUN-REN Hungarian Research Centre for Linguistics
Eötvös Loránd University
juhasz.kornelia8@gmail.com, bartos@nytud.hu

Introduction. Within the scope of the phonology-phonetic interface, discrete and abstract phonological features are presumed to be converted to phonetic targets (Keating 1988). In terms of lexical tones and intonation, phonetic targets are assumed to be realized as turning points in the f0 curve (Keating 1988, Chen & Xu 2006). However, similarly to speech sounds, lexical tones are also affected by the quality of the adjacent tonal patterns, which leads to the formation of contextual tonal variations (i.e. tonal coarticulation) (Xu 1997). In Mandarin Chinese the four lexical full tones phonologically can be characterized by the combination of two underlying targets: high (H) and low (L). High level Tone 1 (T1) features a static H, while low Tone 3 (T3) phonologically features a static L target, but its phonetic realization is mostly described as a mid fall-rise pattern, where the rising phase might be truncated. Rising Tone 2 (T2) and falling Tone 4 (T4) features LH and HL underlying tones, respectively (Xu & Wang 2001). Concerning the directionality of tonal coarticulation in Mandarin, carry-over coarticulatory effects are found to exert a significant (assimilatory) effect on the formation of the subsequent tonal realizations, contrastively to anticipatory effects, which are, although often present (Shen 1990), yet show much weaker (dissimilatory) influence on the preceding lexical tone (Xu 1997). If a carry-over effect is exerted between two adjacent lexical tones, then the low offset of the 1st lexical tone in the sequence – in an assimilatory manner – lowers the onset of the subsequent tone; likewise a high offset of the 1st tone elevates the subsequent tone's onset (Xu 1997). Since anticipatory effects are less salient in the formation of contextual tonal variations, thus in this study we primarily focus on progressive carry-over effects, yet our results include the analysis of anticipatory effects as well. In particular, this preliminary acoustic study focuses on how carry-over tonal coarticulation surfaces in the production of Hungarian learners of Mandarin by analysing utterance-initial trisyllabic lexical tone sequences in declarative sentences, where each syllable appears as an individual morpheme. As Hungarian is an atonal L1, we hypothesize that concatenating lexical tones to sequences poses problems for Hungarian L2 learners, since the sequencing procedure requires the covariation of several factors, such as, tonal context, intonation and focus, as well.

Method. We analysed the production of two groups: intermediate L2 learners' patterns compared to a native MC-speaking group (5 women per group, 10 speakers in total). The recorded trisyllabic tonal sequences were positioned utterance-initially, serving as SVO in broad focus declarative sentences, and were followed by two phonologically unspecified, weak syllables (neutral tones). All combinations of the four MC lexical tones (T1, T2, T3, T4) occurred in the 2nd and the 3rd syllables, while the 1st syllable was fixed high level tone, in this manner we could analyse 16 different tonal sequences. In our analysis the main focus was primarily placed on carry-over tonal coarticulation triggered by the 2nd syllable, affecting the 3rd syllable's realization (in which case all tonal combination appears), however the 1st syllable also exerts carry-over effect on the 2nd syllable, but in this case the analysed tonal combinations are more limited. Since the recorded utterances exclusively consisted of sonorants, f0 could be extracted throughout the trisyllabic sequence automatically by 5 ms intervals in Praat (Boersma & Weeninck 2022). The extracted f0 values were converted to semitones with a reference value of 50 Hz. F0 contours were normalized by their duration syllable-wise, in this manner tonal sequences could be compared by GAMMs (Wood 2017). In the GAMMs the f0 change was analysed dependent on the syllables' normalized duration, and the model was complemented by a parametric term coding the speaker group and tonal value of the 2nd syllable. Beforehand, data was divided by the lexical tone in the 3rd syllable, in this manner we composed four GAMMs in total.

Results. Regarding native Mandarin speakers' production it is apparent how the modification of the lexical tone value in the 2nd syllable significantly influences the realization of the fixed tone in the 3rd syllable, as a result of carry-over coarticulation and also the declarative sentence type (see each column on Figure 1., respectively). Contrastively, in L2 learners' production f0 curves in the 3rd syllable exhibited remarkable similarity, irrespective of the 3rd syllables' tonal value or the preceding 2nd tone. Moreover, L2 learners' f0 patterns were also characterized by an overall compressed f0 range throughout the trisyllabic sequence, compared to the native f0 curves. Turning to the relative temporal positions of f0 inflection points in the 2nd syllable (if existed), though, were realized similarly to natives' production. Although the excursions were less apparent in L2 learners' patterns owing to the compressed f0 range, the significant discrimination of different tonal realizations positioned to the 2nd syllable was still present, approximating native patterns. This also means that the carry-over effect triggered by the utterance-initial T1's static high level target affected both groups' production in a similar manner: the first target's approximation in the 2nd syllable was delayed to the (second) half of the syllables' normalized duration (as was observed by Xu (1997), as well).

Discussion. This acoustic study aimed to shed light on how tonal coarticulation surfaces in the production of Hungarian learners of Mandarin by analysing trisyllabic tone sequences. Our results showed that L2 learners in general differed from native patterns, due to the fact that L2 learners failed to reproduce the carry-over effect on the 3rd syllable triggered by the 2nd syllable of the sequence. In contrast, L2 learners could approximate more the native production in terms of the temporal alignment of the approximation of the 2nd syllable's first target, thus in this acoustic aspect – considering the 1st syllable carry-over effect – L2 learners could produce similar patterns as natives. In sum, we can conclude that based on our results, the production of lexical tones and the concatenation of tone sequences poses problems to Hungarian leaners of Mandarin. In particular, our results suggest that the position of the lexical tone within the trisyllabic sequence influenced L2 learners' production: synchronizing the 1st and the 2nd syllable posed less problems (since in this case the sufficient differentiation between tones in the 2nd syllable was present, even if the patterns did not approximate the native production), in comparison to the interaction of the 2nd and the 3rd syllable (where L2 learners failed to differentiate between the 3rd syllable's different tonal patterns). The results shed light on fundamental problems in lexical tone production and tone sequencing in the case of Hungarian learners of Chinese, and also contribute to the deeper understanding of how tonal coarticulation occurs and tonal contexts interact in the production of atonal learners of Mandarin.

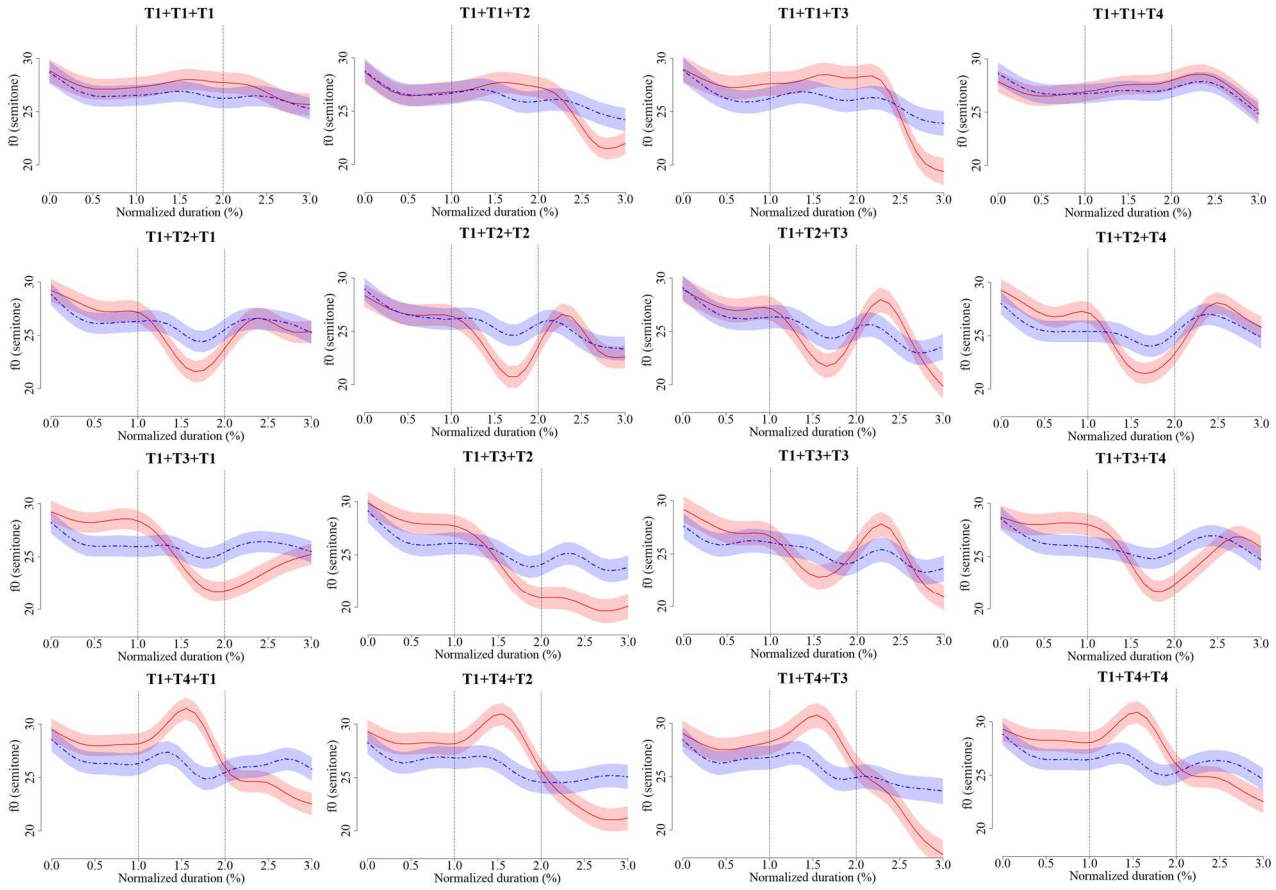


Figure 1: *The f0 curves of the different trisyllabic tonal sequences, where column-wise the 3rd syllable, row-wise the 2nd syllable features identical tonal value, and solid red line represents native, while dashed blue line represents L2 learners' estimated f0 pattern.*

References

Boersma, P. & Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3. http://www.praat.org/

Chen, Y., & Xu, Y. (2006).Production of Weak Elements in Speech – Evidence from F0 Patterns of Neutral Tone in Standard Chinese. *Phonetica, 63,* 47-75.

Keating, P. A. (1988). The phonology-phonetics interface. In Newmeyer (Ed.), *Linguistics: the Cambridge survey.* Cambridge: Cambridge University Press. 281-302.

Shen X. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics, 18*, 281-295.

Wood, S. (2017). *Generalized Additive Models – An Introduction with R*. Boca Ranton: Chapman & Hall.

Xu, Y. & Wang, E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication 33,* 319-337.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*, 61-83.

A constriction geometry analysis of place contrasts in Malayalam nasals

***Alexei Kochetov*[1,2], *Pierre Badin*[2]**

[1]Dept. of Linguistics, Univ. of Toronto, Toronto, Canada
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France
al.kochetov@utoronto.ca, Pierre.Badin@gipsa-lab.grenoble-inp.fr

Introduction. Malayalam (Dravidian) exhibits a typologically unusual 6-way place of articulation contrast in lingual nasal consonants (Kumari, 1972; Namboodiripad & Garellek, 2017), as illustrated in Table 1. How exactly this complex set of contrasts is distinguished by speakers, however, is unclear. The only previous articulatory investigation of a subset of these consonants (/n̪, n, ɳ, ɲ/) by Dart & Nihalani (1999) concluded – based on static palatograms and linguograms from nine speakers – that both the location and the spatial extent of the constriction are important for the characterization of these articulations. In this study we examine the constriction geometry of Malayalam nasals using static MRI data, expanding on the tongue tip constriction angle method proposed by Proctor et al. (2010) designed to model a 4-way coronal contrast in Wubuy stops.

Table 1: *Place contrasts in Malayalam lingual nasals.*

dental	**alveolar**	**retroflex**	**alveolopalatal**	**palatalized (fronted) velar**	**(plain) velar**
pan̪:i	kan:i	kaɳ:i	kaɲ:i	mat:aŋʲ:a	taŋ:i
pig	*a month*	*link*	*gruel*	*pumpkin*	*held fast*

Methods. Single slice mid-sagittal MRI static images were recorded for two native speakers of Malayalam (BB, male; SV, female; both from Thiruvananthapuram, Kerala, India) with a Philips Achieva 3.0T dStream scanner using a 20-channel head-neck coil in Turbo Spin Echo mode. The speakers were asked to produce the nasals /n̪, n, ɳ, ɲ, ŋʲ, ŋ/ (sustaining the articulation for about 6.5 seconds) in 5 symmetric V_V contexts (/a i u e o/, e.g., [aɳa], [iɳi], [uɳu], [eɳe], [oɳo]), as part of a larger corpus of Malayalam sounds. Semi-automatic segmentation of the main speech articulators from the MRI images was performed according to Labrunie et al. (2018). The contours were aligned with the hard palate and two variables were calculated (as in Kochetov et al., 2023): Tongue Constriction *Location* (TCL) and *Length* (TClength). An acoustic Low Frequency Impedance approximation (LFI) was computed for each VT tube as its length divided by the square of its cross-sectional distance. The center of the constriction was considered as the location upstream and downstream of which the cumulated LFIs are equal; TCL was expressed as the angle of this point in reference to the VT center. TClength was estimated as the length of a uniform tube with the same cumulated LFI as the tubes close to the constriction center having a cross-dimensional distance below a given threshold. The results are illustrated in Figure 1, where the constriction limits are outlined by thicker cyan lines on the inner and outer walls, and the center of the constriction is marked by the radial line.

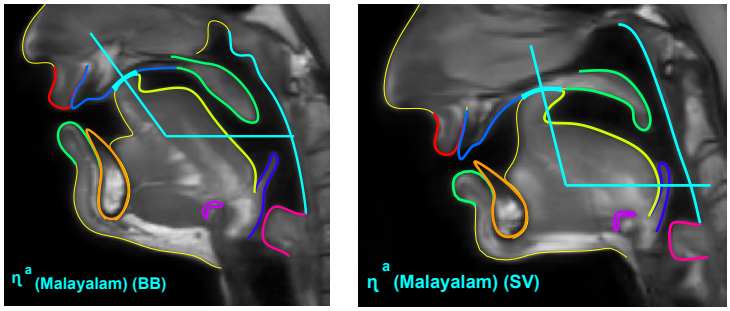


Figure 1: *Articulator contours superimposed on a midsagittal image of /ɳ/ in /aɳa/ by speakers BB and SV with the angle representing the constriction location measure.*

Results. Figure 2 illustrates the tongue constriction location angle (in blue) and constriction length (in green) for all nasal consonants produced by speaker BB in the context /o_o/. It can be seen that the angle progressively decreases from the dental place to the velar place; the constriction length is relatively small for the anterior consonants produced with the tongue tip, blade, or the underside, and is much larger for the posterior consonants produced with the tongue front/body or dorsum. The actual realization of the first three consonants by the speaker can be described as a laminal dental, apical

alveolar, and a subapical palatal retroflex. The last consonant is a fairly posterior velar or uvular, while both /ɲ/ and /ŋʲ/ are laminal alveolopalatals yet slightly different in the frontness of the constriction (and the overall advancement of the tongue). Results of Linear Mixed Effects Regression (LMER) models and posthoc tests performed separately by variable and speaker (see Table 2) revealed that TCL angle distinguished almost all places for speaker BB (with the exception of posterior coronals - retroflex and alveolopalatal) and a subset of places for speaker SV: anterior coronals and plain velar from posterior coronals and palatalized velar. For both speakers, TC length distinguished dentals, alveolars, and retroflexes from alveolopalatals and the two velars. Taken together, all six place contrasts were distinguished by the two measures for BB, while all but the dental and alveolar (/n̪, n/) were distinguished by SV.

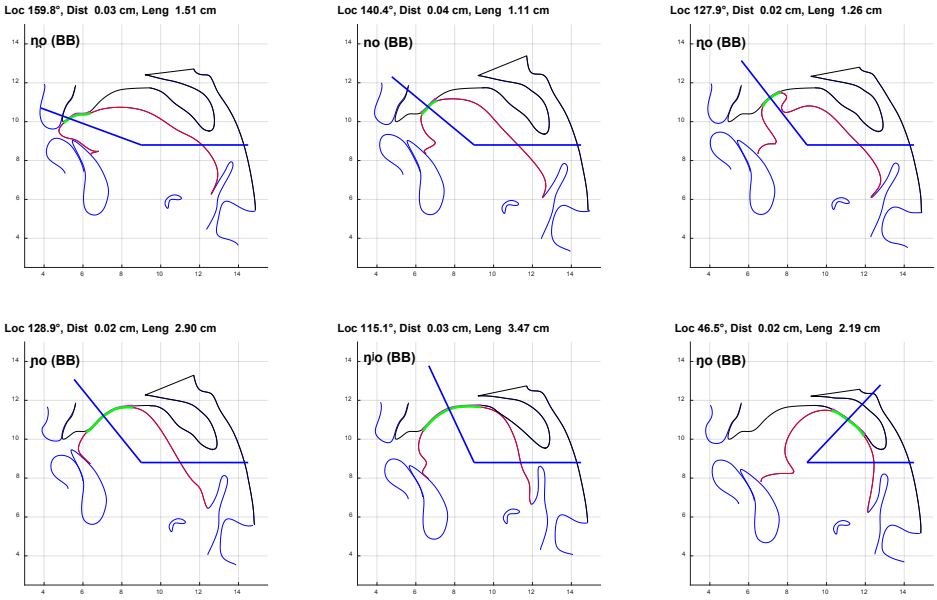


Figure 2: *Constriction location plots for nasal consonants in the /o/ context by speaker BB.*

Table 2: *Results of LMER model comparisons for Tongue Tip Constriction Location (TTCL) and Constriction Length (TTlength) and pairwise posthoc comparisons by speaker.*

Speaker	Variable	DF	F	Pr(>F)	Posthoc differences
BB	TTCL	5	41.25	<.001	n̪ > n > ɳ, ɲ > ŋʲ > ŋ
	TTClength	5	40.38	<.001	ɲ, ŋʲ, ŋ > n̪, n, ɳ
SV	TTCL	5	18.24	<.001	n̪, n > ɳ, ɲ, ŋʲ > ŋ
	TTClength	5	12.77	<.001	ɲ, ŋʲ, ŋ > n̪, n, ɳ

Discussion. The results showed that measures of TCL angle and length are useful in distinguishing a complex set of lingual place contrasts in Malayalam, consistently with Proctor et al.'s (2010) work on Wubuy coronal stops and Kochetov et al.'s (2023) analysis of Kannada dentals and retroflexes. As the results for speaker SV showed, however, TCL measures may not be always sufficient for distinguishing dentals and alveolars, and thus may need to be complemented by articulatory modeling components such as Tongue Tip Fronting and Raising. We are currently exploring this approach.

References

Dart, S.N., & Nihalani, P. (1999). The articulation of Malayalam coronal stops and nasals. *Journal of the International Phonetic Association*, *29*, 129-142.

Kochetov, A., Savariaux, C., Lamalle, L., Noûs, C., & Badin, P. (2023). An MRI-based articulatory analysis of the Kannada dental-retroflex contrast. *Journal of the International Phonetic Association*, 1-37.

Kumari, S.B. (1972). *Malayalam phonetic reader*. Mysore, India: Central Institute of Indian Languages.

Labrunie, M., Badin, P., Voit, D., Joseph, A.A., Frahm, J., Lamalle, L., Vilain, C., & Boë, L.-J. (2018). Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. *Speech Communication*, *99*, 27-46.

Namboodiripad, S., & Garellek, M. (2017). Malayalam (Namboodiri Dialect). *Journal of International Phonetic Association*, *47*, 109-118.

Proctor, M., Bundgaard-Nielsen, R.L., Best, C.T., Goldstein, L., Kroos, C., & Harvey, M. (2010). Articulatory modelling of coronal stop contrasts in Wubuy. In *SST 2010, 13th Australasian Speech Science and Technology*, pp. 90-93. Melbourne, Australia.

Effect of varying rhythmic stimulations on fluency and production gestures of people who stutter

Maëva Garnier, Anneke Slis, Victor Juliard, Christophe Savariaux

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

maeva.garnier@gipsa-lab.fr

Introduction. Speaking with a metronome has a facilitating effect on the speech of people who stutter (PWS) (Brady 1969), reducing the frequency and severity of disfluencies (Hanna & Morris 1977; Kalinowski et al., 2000). A number of studies have shown that this perceptual improvement is accompanied by objective changes in speech production, both acoustically (Brayton & Conture, 1978; Klich & May, 1982; Stager et al., 1997; Davidow et al. 2011) and aerodynamically (Hutchinson & Navarre, 1977; Stager et al., 1997). In a recent study, we also showed that speaking with metronome contributes to reduced levels of lip muscle activity during labial stop consonants, as well as to shorter voice onset times, which both tended to be greater in the natural speech of PWS (Garnier et al. 2023).
The goal of this study is to compare this facilitating effect for varying types of rhythmic stimulations, in order to better understand *why* and *how* speaking with a metronome improves fluency. The question is whether this simply leads speakers to slow down their speech rate, whether this helps them replace the complex prosodic pattern by a simpler rhythmic pattern, or whether providing external triggers helps them initiate motor sequences (Alm, 2004).

Methods. Sixteen French adults who stutter (PWS) were matched in age, gender and musical experience with sixteen normofluent adults (PNS). In a first reference task (REF), they produced, 20 sentences of 7 syllables, containing bi-syllabic words beginning with /pa/ or /ba/, at a comfortable rate, without any rhythmic constraint (e.g. "Pattie passa la pagaie"). The same sentences were used in 5 rhythmic conditions during which the participant synchronized each syllable with a metronome beat: 1- normal paced at 120 BPM ($SYNC_{120}$), 2- fast paced at 240 BPM ($SYNC_{240}$), 3&4- complex paced, with two different non-isochronous kind of "musical" patterns ($NONISO_1$, $NONISO_2$), and 5- producing each syllable as soon as possible after hearing aperiodic stimuli (REACT). The audio signal of these productions was recorded simultaneously with the electromyographic (surface) activity of the orbicularis oris superior (OOS) and inferior (OOI).
Disfluencies were perceptually detected from the audio signal by a certified speech therapist and the percentage of disfluent syllables was calculated for each speaker and each condition. Only for those not annotated as disfluent, the energy of both EMG signals was calculated over the duration of the syllables /pa/ and /ba/. Statistical analyses were conducted from mixed models of the data, considering the 6 conditions and the group (PNS vs. PWS) as fixed effects, and the consonant voicing (/p/ vs. /b/) and the participant as random intercepts.

Results. On the perceptual level, all 5 rhythmic conditions led to significant reduction in the frequency of disfluencies, compared to the reference non-rhythmic task. The greatest reduction was observed for the $SYNC_{120}$ condition, in which almost no disfluency was observed, followed by the $SYNC_{240}$ one, then the two non-isochronous conditions $NONISO_1$ and $NONISO_2$, and finally the REACT condition (see top panel of Figure 1). In terms of production, the significant difference in VOT and lip muscle activity, observed between the two groups in the reference task (REF), was no longer observed in the metronome condition $SYNC_{120}$ (see bottom 2 panels of Figure1). A smaller difference was still observed in all the rhythmic tasks, remaining significant for the conditions $NONISO_1$, $NONISO_2$ and REACT, but small enough to become non-significant for the faster metronome condition $SYNC_{240}$.

Discussion. First, our results show, like in previous work, that the speech gestures of PWS demonstrate significant atypicality, compared to productions of PNS, even outside episodes of disfluency, in agreement with the idea that the categorical perception of speech as fluent or disfluent may actually be underlined by a physiological continuum of atypicality in production gestures (Hulstijn & Van Lieshout, 1998). Further supporting this idea, the comparison of speech produced in the different rhythmic conditions shows that the frequency of disfluencies in PWS can be relatively well predicted by the size and significance of acoustic and physiological differences between their productions and those of the typical group. However, our results do not show a simple relationship between the variation of acoustic or physiological parameters, and the frequency of disfluencies: thus, the fluency improvement cannot be simply related to a shortening of the VOT, or to a reduced lip muscle activity (see bottom 2 panels of Figure1).
Furthermore, our study provides quantitative evidence that metronome paced speech improves fluency in PWS at the fine-grained articulatory level. The comparison of the different rhythmic conditions provides new insights into their facilitating effect and possible neurological deficits in PWS, which these rhythmic stimulations may compensate for. Thus, the almost comparable fluency improvement and reduced atypicality of speech gestures in the fast metronome condition $SYNC_{240}$ as well as in the slower one $SYNC_{120}$, rules out the hypothesis that the beneficial effect of the metronome simply comes from slowing down the speech rate. The significant improvement in fluency that is still

observed in the REACT condition – although to a lesser extent than in the metronome conditions ($SYNC_{120}$ and $SYNC_{240}$) – remains compatible with the hypothesis that rhythmic stimulations, regardless of their complexity and predictability, may help initiate motor sequences, by providing external triggers in place of possibly deficient internal triggers, in relationship to the dysfunctional cortico-basal ganglia-thalamo-cortical loop in PWS (Alm, 2004). Finally, the significant improvement of speech fluency in the non-isochronous conditions $NONISO_1$ and $NONISO_2$, to an intermediate level between that observed in the reference condition (low), and that observed in the metronome $SYNC_{120}$ condition (high), also supports the hypothesis that rhythmic stimulations may facilitate speech planning, by replacing the complex prosodic pattern of a utterance by a simpler rhythmic pattern. These various hypotheses deserve to be explored in greater detail in future work. These results also raise the question of how to use such rhythmic stimulations outside of the speech therapist's office, in real situation of communication. In addition, the question is also if the reduced atypicality of speech gestures that these stimulations induce can then be learnt and transferred more durably, after the stimulation stops.

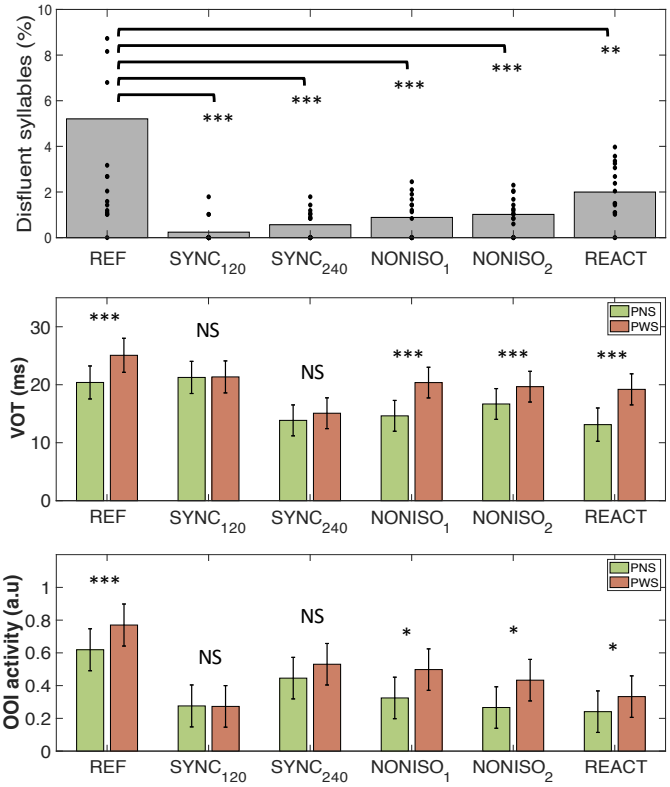


Figure 1: *Top panel: Frequency of disfluent syllables, for people who stutter only, in the non-rhythmic task of reference, and in the 5 varying rhythmic tasks. Bottom two panels: Average Voice Onset Time (VOT) and EMG activity of the Orbicularis Oris muscle (OOI) during the production of fluent /pa/ and /ba/ syllables by people who stutter (PWS) and people who not stutter (PNS), for the same 6 conditions.*

References

Alm, P. A. (2004). Stuttering and the basal ganglia circuits: a critical review of possible relations. Journal of communication disorders, 37(4), 325-369.

Brady, J. P. (1969). Studies on the metronome effect on stuttering. Behaviour Research and Therapy, 7(2), 197-204.

Brayton, E. R., & Conture, E. G. (1978). Effects of noise and rhythmic stimulation on the speech of stutterers. Journal of Speech and Hearing Research, 21(2), 285-294.

Davidow, J. H., Bothe, A. K., & Ye, J. (2011). Systematic studies of modified vocalization: speech production changes during a variation of metronomic speech in persons who do and do not stutter. Journal of fluency disorders, 36(2), 93-109.

Hanna, R., & Morris, S. (1977). Stuttering, speech rate, and the metronome effect. Perceptual and Motor Skills, 44(2), 452-454.

Hutchinson, J. M., & Navarre, B. M. (1977). The effect of metronome pacing on selected aerodynamic patterns of stuttered speech: Some preliminary observations and interpretations. Journal of Fluency Disorders, 2(3), 189-204

Hulstijn, W., & Van Lieshout, P. H. H. M. (1998). A motor skill approach to stuttering. Clinical phonetics and linguistics, 391-404.

Kalinowski, J., Stuart, A., Rastatter, M. P., Snyder, G., & Dayalu, V. (2000). Inducement of fluent speech in persons who stutter via visual choral speech. Neuroscience letters, 281(2-3), 198-200.

Klich, R. J., & May, G. M. (1982). Spectrographic study of vowels in stutterers' fluent speech. Journal of Speech, Language, and Hearing Research, 25(3), 364-370.

Stager, S. V., Denman, D. W., & Ludlow, C. L. (1997). Modifications in aerodynamic variables by persons who stutter under fluency-evoking conditions. Journal of Speech, Language, and Hearing Research, 40(4), 832-847.

Prenasalization in initial voiced stops in Zuberoan Basque

Ander Egurtzegi,[1,2] Iñigo Urrestarazu-Porta[1,2,3,4]&Andrea García-Covelo[5,2,3]

[1]*Centre national de la recherche scientifique (CNRS);* [2]*IKER-UMR5478;* [3]*University of Pau (UPPA);*
[4]*University of the Basque Country (UPV/EHU) &* [5]*Institute for Phonetics and Speech Processing (IPS), LMU Munich*
ander.egurtzegi@iker.cnrs.fr,
inigo.urrestarazu-porta@iker.cnrs.fr & andrea.garcia@phonetik.uni-muenchen.de

Introduction. The aerodynamic voicing constraint or AVC (Ohala 1983; Ohala 2011) refers to the requirement that sufficient airflow has to pass through the adducted vocal folds for voicing to be maintained during the production of a voiced obstruent or, in the specific case discussed in this paper, during the closure phase of a voiced oral stop. To ensure this, subglottal pressure has to be sufficiently higher than supraglottal pressure, a situation unlikely to be sustained due to the accumulation of air in the oral cavity during the stop closure. Vocal fold vibration stops when the pressure difference is not large enough. Thus, the AVC greatly reduces the time during which voicing can be maintained in voiced oral stops as opposed to voiced sonorants. There are some strategies that are typically implemented in order to maintain voicing in voiced oral stops for longer periods of time, such as passive vocal tract enlargement (Ohala and Riordan 1979). Nonetheless, in many cases, the AVC will result in sound change. Not preventing it would result in stop devoicing (especially in posterior stops), and avoiding it might result in a variety of sound changes including spirantization, prenasalization, the development of implosives or [ATR], and the retraction of apicals (Ohala 2011).
In Basque, word-medial voiced stops have been often reported to be produced as approximants (Hualde 2003), with a phonological distribution similar to that described for Spanish. While it is known that word-initial (or, more precisely, utterance-initial) voiced oral stops do not necessarily show this lenitive process, they should also be subject to the AVC, as observed in Spanish (Solé and Sprouse 2011). This paper investigates the possibility that the AVC in utterance initial stops results in voiced stop prenasalization in Basque with data from the Zuberoan variety.

Methods. Our recordings were made in the Zuberoan village of Larraine. Local participants performed a reading task where isolated words were elicited. We recorded the utterances of 6 volunteer native speakers of Zuberoan Basque (5 male, 1 female; mean age 65, range 60-70) using a nasalance device with a separator handle, which consists of two microphones separated by a wooden plate that facilitates the separation of the acoustic signal coming from the mouth and that coming from the nose. The stimuli were randomized, presented and recorded with the *SpeechRecorder* software.
The recordings were originally meant to measure nasality in aspirates (Egurtzegi, García-Covelo, and Urrestarazu-Porta 2023), but they included 166 tokens with an initial voiced stop, which could be followed by any of the 6 vowels in Zuberoan Basque (/a, e, o, i, y, u/). In total, there are 124 tokens with word-initial /b/ (e.g. *behi* 'cow'), 25 with /g/ (e.g. *gehien* 'most'), and 17 with /d/ (e.g. *dohan* 'free', note that initial /d-/ is rare in the language). We included initial voiceless stops, vowels and nasal stops as control conditions (see Figure 1). The stereo nasalance data was processed using Praat. For the acoustic analysis, both the nasal and oral channels were band-pass filtered (80 Hz-10000 Hz) and the nasalance (ratio of the nasal amplitude to the sum of the oral and nasal amplitudes, i.e. $A_n/(A_o + A_n) \times 100$) (Carignan 2018) was computed every 5 ms. We then obtained the median, minimum and maximum nasalance values of each production.
Three Bayesian regression models with the same structure were fitted: Median, minimum or maximum nasalance values as dependent variables, category as the independent variable, by-speaker correlated varying slope and intercept adjustments and by-word varying intercept adjustments. We included weakly informative priors to stabilize the posteriors.

Results. The model on median nasalance values shows a clear distinction between all four groups, voiced stops having median nasalance values between vowels and nasal stops. Regarding nasalance minima, the model estimates voiced stops to have greater nasalance values than voiceless stops and lower values than nasal stops, while vowels and voiced stops cannot be distinguished. The model estimates on nasalance maxima show similar values for voiced stops and nasals. Vowels and voiceless stops have a lower estimated nasalance maximum and cannot be distinguished from each other.

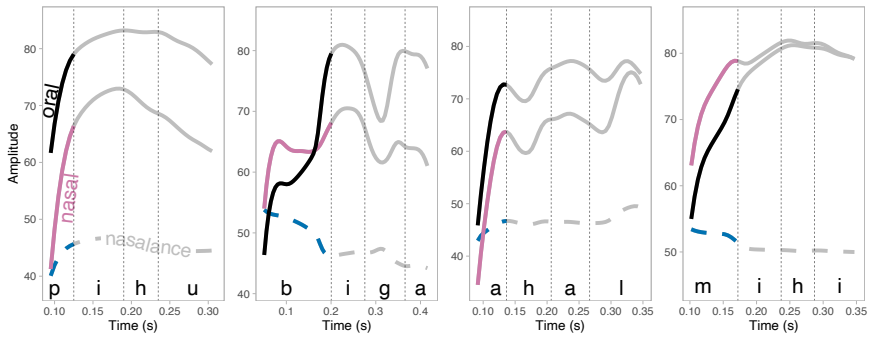


Figure 1: *Amplitude of oral and nasal channels and nasalance across time. In voiced stops (second) and nasals (fourth) the amplitude of the nasal channel is above the amplitude of the oral channel and nasalance values are higher.*

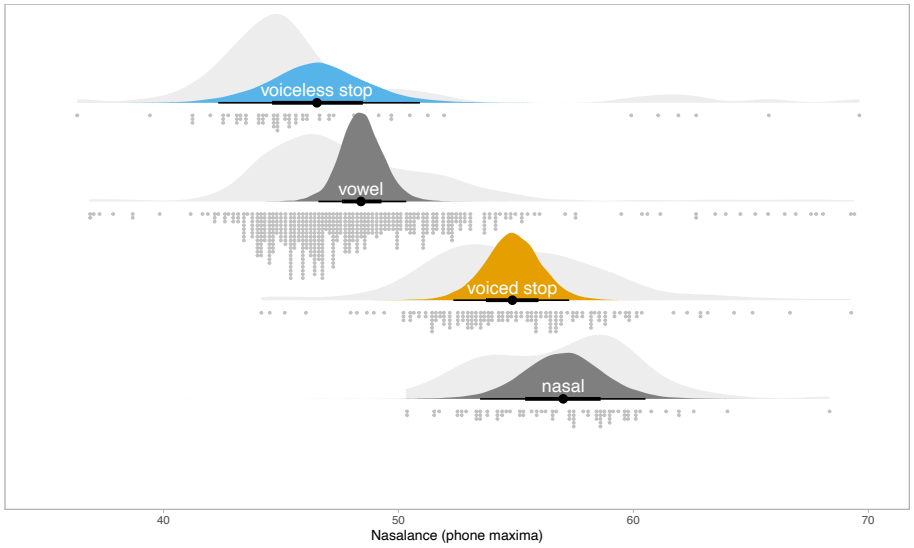


Figure 2: *Nasalance maxima by category. The distributions indicate posterior distributions of the model. The dots are a histogram of the data. Shades in the back are density distributions of the data.*

Discussion. Nasalance maxima are similar for voiced oral stops and nasals. This indicates that voiced stops are nasalized at some point of their duration, which should be interpreted as a peak in nasality. If voiced stops were fully nasalized we would expect them to have similar median and minimum nasalance estimates as nasals. Yet, median nasalance estimates of voiced stops lay between those of vowels and nasals and estimated nasalance minima fail to distinguish between vowels and voiced stops. Finally, visual inspection of the amplitudes of the oral and nasal channels indicate that the peak of nasality happens at the beginning of the segments (see Figure 1), which suggests voiced stops are in fact prenasalized.

References.

Carignan, C. (2018). "Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels". In: *The Journal of the Acoustical Society of America* 143 (5), pp. 2588–2601.

Egurtzegi, A., A. García-Covelo, and I. Urrestarazu-Porta (2023). "A nasalance-based study of the /h/ vs. /ˉh/ Ooposition in Zuberoan Basque". In: *ICPhS XX*. Ed. by R. Skarnitzl and J. Volín. Prague: Guarant International, pp. 3427–3431.

Hualde, J.I. (2003). "Segmental phonology". In: *A grammar of Basque*. Ed. by J.I. Hualde and J. Ortiz de Urbina. Berlin: Mouton de Gruyter, pp. 16–65.

Ohala, J.J. (1983). "The origin of sound patterns in vocal tract constraints". In: *The Production of Speech*. Ed. by P.F. MacNeilage. New York: Springer-Verlag, pp. 189–216.

— (2011). "Accommodation to the aerodynamic voicing constraint and its phonological relevance". In: *ICPhS XVII*. Hong Kong: IPA, pp. 64–67.

Ohala, J.J. and C.J. Riordan (1979). "Passive vocal tract enlargement during voiced stops". In: *Speech communication papers*. Ed. by J.J. Wold and D.H. Klatt. New York: Acoust. Soc. of Am., pp. 89–92.

Solé, M.J. and R.L. Sprouse (2011). "Voice-initiating Gestures in Spanish: Prenasalization". In: *ICPhS XVII*. Hong Kong: IPA, pp. 72–75.

Vocal expression of emotions by patients with Unilateral Vocal Fold Paralysis

Caterina Petrone[1], Nicolas Audibert[2], Ralph Haddad[3], Méline Robert[1], Marion Trocq[1], Alexia Mattei[1,3], Muriel Lalain[1]

[1] *Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France*
[2] *Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne Nouvelle, France*
[3] *Hôpital La Conception, Marseille, France*

caterina.petrone@univ-amu.fr, nicolas.audibert@sorbonne-nouvelle.fr, ralph.haddad@ap-hm.fr, meline.robert24@gmail.com, marion.trocq@hotmail.fr, alexia.mattei@ap-hm.fr, muriel.lalain@univ-amu.fr

Introduction. Prosody provides a powerful means to express basic emotions (e.g., anger or sadness) and, as such, it contributes to effective communication and social functioning (Scherer, 2003). High arousal emotions like (hot) anger are characterized in many languages by an increase in the fundamental frequency (f0) mean, higher intensity, and harsh/tense voice compared to a neutral emotional state (e.g., Scherer, 2003; Gobl and Ní Chasaide, 2003). Low arousal emotions like sadness are often characterized by a decrease in f0 and intensity and by an increase in breathiness (Scherer, 2003; Gobl and Ní Chasaide, 2003). While, in healthy adults, basic emotions are associated to systematic prosodic modulations, patients with unilateral vocal fold paralysis (UVFP) complain of a mismatch between the emotion they intend to express and the emotion effectively conveyed through their voice (Mattei, p.c.). This exploratory study aims at providing a first assessment of the impact of UVFP in the vocal expression of emotions. UVFP consists in an immobility of one of the vocal folds, arising from an interruption of motor nerve control of the intrinsic muscles of the larynx (Alwan & Paddle, 2021). Typical symptoms of UVFP include dysphonia and instability in the vibratory pattern of the vocal folds. Such symptoms may lead to compensatory adjustments further increasing patients' vocal effort (Manal & Gamal, 2015). UVFP patients report weak voice, breathiness, roughness, diminished voice intensity, diplophonia, air loss (Lotto et al., 1997; Jesus et al., 2015). Acoustically, UVFP results in higher values of jitter and shimmer, lower values of the harmonics-to-noise ratio (HNR) and lower f0 range compared to heathy controls (Hirano et al., 1995; Jesus et al., 2015). We expected that the global increase in breathiness and roughness and the decrease in f0 control will lead to no or smaller acoustic differences in the expression of sadness, anger and neutral state compared to controls.

Methods. A sample of ten French UVFP patients (mean age: 66, min = 55 y.o., max = 77 y.o; 5 women and 5 men) and ten French control speakers (matched in age and sex) has participated so far in our study. We included only patients with a post-operative UVFP (e.g., paralysis secondary to thyroidectomy, endarterectomy or cardiac surgery), with no dysarthria and no neurological or psychiatric disorders. UVFP patients underwent standard voice assessment, that included a VHI self-questionnaire, the Hirano's GRB scale and the measurement of maximum phonation time (see Mattei et al., 2018 and references therein). All participants performed a sentence production task. Materials for this task included eight sentences with verbal neutral meaning, that were validated in a prior study (e.g., *Il va rentrer chez lui*, "He is going back home"). Sentences were short (five to nine syllables) and had the same syntactic structure. The verb was always a periphrastic form of near future (*va* followed by the infinitive of the verb: e.g., *va rentrer* "going back"). Each sentence was embedded in three different contexts, eliciting three different emotional states (neutral/sad/angry). Participants read all the contexts and target sentences silently, and then they produced the target sentences in a natural way without reading. To facilitate the task, sentences were presented in three different blocks of neutral, sad and angry emotional states. The intended emotion was indicated at the beginning of each block, and each block was preceded by a familiarization and a training phase. Within each block, sentences were presented in a random order. We collected 480 utterances (8 sentences X 3 emotions X 10 participants X 2 populations). Acoustic measures were extracted at the midpoint of vowel /a/ of the word *va*, as it occurred in all sentences. The vowel /a/ in each sentence was manually segmented by trained speech scientists. Fundamental frequency (f0) within the /a/ was extracted using FCN-f0 (Ardaillon and Roebel, 2019), evaluated as more reliable on pathological speech compared to other pitch detection algorithms (Vaysse et al., 2022). To account for possible voice quality differences, HNR over 1kHz and CPPS were extracted using a custom Praat script. Moreover, the 0-5kHz spectrum was computed on 20Hz bins. For each measure, a linear mixed model was fitted to evaluate the effect of the speaker group (control speaker vs. patients), the intended emotion (neutral/sad/anger) and their interactions. Spectra of /a/ uttered by the same speaker were compared across intended emotions by computing the correlation coefficient. The significance of the main effects was assessed by comparing the complete model with the model without the effect tested. Post-hoc comparisons were made using estimated marginal means, with p-values adjusted using Tukey's method.

Results and discussion. The comparison of log-transformed vowel durations shows that segmental duration is longer for patients than for control speakers ($\chi^2(1)=10.80$; p=.001). Analysis of CPPS values confirms that patients are more

dysphonic than controls ($\chi^2(1)=21.74$; p<.001), without significant differences between intended emotions in the patients' group. Visual inspection of average spectra (Fig1a) suggests that, while expressions of anger by controls are associated with a distinct spectral shape and higher intensity than sadness and neutral, those of patients are characterized by an increase in intensity only. The analysis of power-transformed correlation coefficients shows a strong interaction between speaker group and emotion ($\chi^2(2)=71.48$; p<.001). Post-hoc comparison confirms that the difference in spectral shape between neutral and sadness is comparable between controls and patients (t(20.8)=-1.19; p=.248), but that the difference between neutral and anger is greater for controls than for patients (t(20.8)=-4.51; p<.001). Fig1b displays the distribution of f0 in sadness and anger relative to each speaker's neutral expression, with a significant interaction between speaker group and emotion ($\chi^2(1)=22.95$; p<.001). Sadness is characterized by f0 values not significantly different from the neutral expression for either patients (t(22.8)=-0.40; p=.691) or controls (t(22.8)=-0.58; p=.565). For anger, f0 values are higher and more distinct from neutral in controls than in patients (t(21.2)=2.61; p=.016). Overall, f0 values have a smaller range of variation in patients for contrasting emotions. HNR values (Fig1c) show a less rich harmonic structure in the patients ($\chi^2(1)=4.88$; p=.027), without significant differences between emotions, suggesting that there are few acoustic differences in the expression of the three emotions sadness, anger and neutral in this group. For the controls, the higher HNR in anger compared to sadness (t(456)=3.14; p=.005) indicates a greater harmonic richness, which helps distinguishing between the three emotional states. Our results support patients' informal observations that UVFP has a negative impact on their ability to convey emotions, possibly because of their reduced possibilities in modulating prosodic features. We further aim at investigating the correlations between voice assessment scores and individual speech performances.

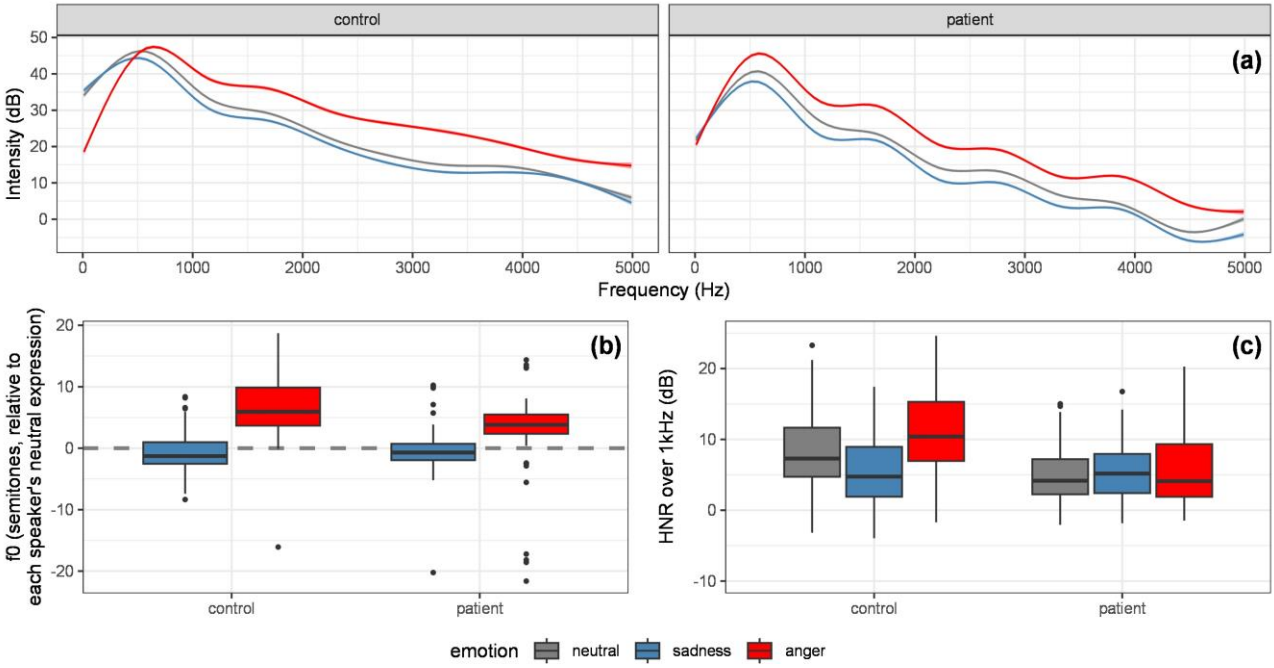


Figure 1. Acoustic comparison of the realization of /a/ on the three intended emotions, for controls and UVFP patients: (a) average spectra from 0 to 5kHz, (b) f0 in semitones compared to the neutral expression produced by the same speaker, (c) harmonic-to-noise ratio (HNR) computed after high-pass filtering to retain only frequencies above 1kHz.

References

Alwan, M., Paddle, P. M. (2021). Vocal Cord Paralysis: Pathophysiology, Etiologies, and Evaluation. *Int. J. Head Neck Surg.*, 12(4), 153–160

Ardaillon, L., Roebel, A. (2019) Fully-Convolutional Network for Pitch Estimation of Speech Signals. *Proc. Interspeech* 2019, 2005–2009.

Gobl, C., and Chasaide, A. N. (2003). The Role of Voice Quality in Communicating Emotion, Mood and Attitude. *Speech Communication*, 40, 189–212.

Jesus, L. M. T. , Martinez, J., Hall, A., Ferreira, A. (2015). Acoustic Correlates of Compensatory Adjustments to the Glottic and Supraglottic Structures in Patients with Unilateral Vocal Fold Paralysis. *Biomed. Res. Int.*, 704121.

Hirano, M., Mori, K., Tanaka, S., Fujita, M. (1995). Vocal function in patients with unilateral vocal fold paralysis before and after silicone injection. *Acta Otolaryngol*, 115(4):553-9.

Lotto, A. J., Holt, L. L., Kluender, K. R. (1997). Effect of Voice Quality on Perceived Height of English Vowels. *Phonetica*, 54(2):76–93.

Manal E.-B., Gamal, Y. (2015). Early Voice Therapy in Patients with Unilateral Vocal Fold Paralysis. *Folia Phoniatr Logop* 1, 66 (6): 237–243.

Mattei, A., Desuter, G., Roux, M., Lee, B.-J., Louges, M.-A., Osipenkof, E. et al. (2018). International consensus (ICON) on basic voice assessment for unilateral vocal fold paralysis. *Eur. Ann. Otorhinolaryngol. Head Neck Dis*., 135, S11–S15.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1–2), 227–256.

Vaysse, R., Astésano, C., & Farinas, J. (2022). Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech. *JASA*, *152*(5), 3091–3101.

Lingual ultrasound feedback in L2 pronunciation practice in classroom: a pilot study of French mid front-back vowel contrast

Claire Pillot-Loiseau[1], Hélène Gustin-Masset[1], Tanja Kocjančič Antolík[2] and Takeki Kamiyama[3]

[1] Laboratoire de Phonétique et Phonologie, UMR 7018 CNRS/Sorbonne-Nouvelle, Paris, France
[2] Faculty of Arts, Charles University, Prague, Czech Republic
[3] Université Paris 8 Vincennes - Saint-Denis, TransCrit, Saint Denis, France
claire.pillot@sorbonne-nouvelle.fr, helene-marie.masset@orange.fr,
tanja.kocjancicantolik@ff.cuni.cz, takeki.kamiyama@univ-paris8.fr

Introduction. Several works have shown the effectiveness of individual pronunciation lessons using lingual ultrasound visual feedback (LU-VF) to help learners better produce L2 sounds: among others, the /y-u/ contrast for Japanese-speaking learners of French (Kocjančič Antolík *et al.* 2019). However, can the same method be applied to classroom teaching? To date, only two pilot studies addressed this issue (Meadows 2007, Kühnert & Pillot-Loiseau 2022), and reported an improvement in the production of some of the trained speech sounds. Neither of them, however, addressed auditory discrimination and identification of the same sounds in French as a foreign language (FFL).
The main aim of the current study was to attest the evolution in L2 vowel production and perception brought by the use of LU-VF in a classroom. The practice was focused on the contrasts /ø-o/ and /œ-ɔ/ learned by L1 Japanese- or Spanish-speaking adult intermediate FFL learners (JSL and SSL, respectively). Given the differences between the 5-vowel systems in their L1 (Vance 2008, Hualde 2005) and French with front rounded vowels and 4 degrees of vowel height (Fougeron and Smith 1993), it can be predicted that these learners will have difficulty perceiving and producing the French mid front rounded vowels /ø/ and /œ/ (Racine and Detey 2018), with JSL assimilating them with /u/ (Kamiyama *et al.* 2017) and SSL, /o/ and /ɔ/ (Racine 2017).
If perception precedes production in L2 (Best & Tyler 2007, among others), we can hypothesize that the learners who have difficulties producing L2 vowels will also have difficulty perceiving them. Based on earlier reports, we can expect that LU-VF will lead to improvements in the production of target vowels, but little is known about its impact on perception. Considering the DIVA model of speech motor control (Guenther 1994), production of a frequent syllable or individual sound starts with a feedforward activation of prestored sensory states and motor plan state. Sensory state includes somatosensory, auditory and visual (external, from other speakers, or internal when observing one's own articulators) information (Kröger & Kannampuzha 2008). This activated sensory state serves as a reference point for the sensory feedback available once the production is complete. When practicing L2 vowel production with LU-VF, the learner must first generate a new visual sensory state. Once the visual feedback serves as a decision marker for correctness in repeated practice, the resulting new somatosensory and auditory feedback and motor plan are created: the learner learns to produce correct targets and to perceive them as correct.

Methods. This study concerns the production and perception of /ø/-/o/ and /œ/-/ɔ/ contrasts by 7 native speakers of French, and two groups of adult intermediate FFL learners (24-48 y.o., living in France for 6 months - 3 years): a control (CTR: 6 JSL and 2 SSL) and an experimental groups (EXP: 2 JSL, 1 SSL). All learners attended an in-class group course in French pronunciation (12 lessons over 4 months) with the same teacher. During the class, EXP received 10 minutes of additional practice in front of the class with LU-VF in 7 of the lessons, in which they observed their tongue while producing words in isolation containing the target vowels and comparing their own LU-VF images with those of a model native speaker. The other students were instructed to listen and to watch the LU-VF images projected on their computer screens in comparison with the image of their classmates' production, and to express an opinion on its accuracy or otherwise. All participants were audio-recorded reading (i) French /u, ø, o, y, i, a/ in isolation (V-isol: 5 repetitions), (ii) 8 carrier sentences containing the vowels in phonotactically plausible contexts, namely, /y, u, ø, o/ in /pVs/ /pVz/ and /œ, ɔ/ in /pVʁ/ (8 repetitions), (iii) a text containing the target vowels in various segmental and prosodic contexts. The learners were recorded before (T1) and just after (T2) the four-month pronunciation course. Additionally, they underwent pre- and post-training perceptual identification test of /y, u, ø, o, œ, ɔ/ in the same words as used in the production task.

Results. Formant analysis showed that, although some learners marked each contrast in one way or another in their production before the training, the two contrasts were most frequently marked by different non-native-like realizations: /ø/-/o/ by too low F2 for /ø/, /œ/-/ɔ/ by too low F1 for both vowels and too low F2 for /œ/. After the practice, an increase in F2 was observed for /ø/ (Figure 1, left), /œ/ (Figure 1, right) and /ɔ/, but only in EXP. The evolution was also more

notable for learners who had been speaking French for a shorter time. The difference between the contrasted vowels was greater for text reading than monosyllables in carrier sentences for all native speakers and learners.
Most learners performed near ceiling (two at 100%) in the forced-choice identification test between the two vowels of each contrast, and both groups showed additional gains after training. No significant difference was observed between EXP and CTR: at T1, CTR showed a mean correct identification rate (all vowels included) of 90.6%, and 94.2% at T2; EXP showed 90.6% at T1 and 95% at T2.

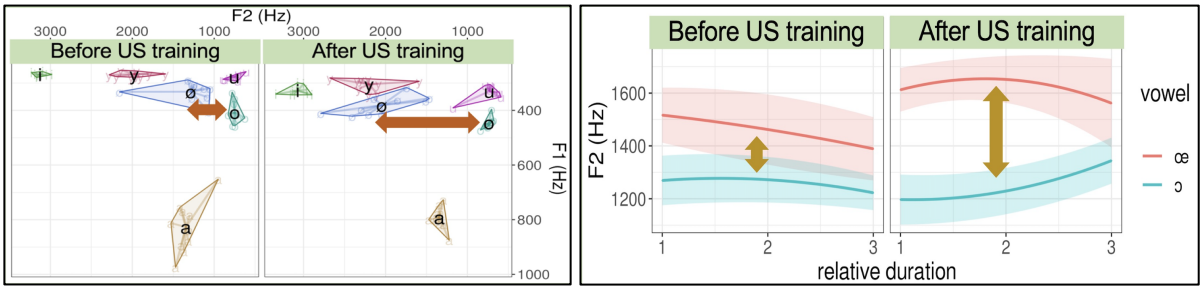


Figure 1: *Left: vowel space of a JSL (vowels in isolation) in EXP. Right: F1 and F2 of vowels in sentences for a SSL in EXP: before and after LU-VF training. Normalized data using z-scores will be presented at the conference.*

Discussion. Despite considerable variability in learner recruitment and responses, as any teacher may encounter in a language classroom, the study confirms the feasibility of LU-VF in an L2 classroom. The results show that although all learners perceptually identified the vowels in /ø/-/o/ and /œ/-/ɔ/ contrasts correctly, they all had more difficulty in producing them. After the LU-VF training, some acoustic changes, mainly those which might stem from horizontal tongue positioning, were observed: the Euclidean distance between /ø/ and /o/ in EXP increases only for V-isol, as F2 of /ø/ increases after the LU-VF lessons; F2 of /ø/ sometimes gets close to F2 of /y/ or /u/ (Racine and Detey 2018; Kamiyama and Vaissière 2009). The effect of LU-VF on perception, and the link between L2 perception and production are less clear. The learners did not have difficulties in L2 perception, likely resulting from the L1-L2 sound merge (Best & Tyler 2007), which in turn affects production. Difficulties in production can be explained by the DIVA model, since speakers with wider regions in the sensory space (resulting from merged L1 and L2 categories) should be less precise in the production of individual sounds (Li *et al.* 2019). To make a permanent change, the new sound has to be fully acquired so that only this "correct" version of the sensory, and motor plan states are activated in production. The amount of practice in the current study was most likely not enough to cause such permanent change. Any future studies should additionally include perceptual assessment of the vowel contrast produced by learners.

References

Best, C. T., & Tyler, M. D. (2007) Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds), *Second Language speech learning: the role of language experience in speech perception and production* (pp.13–34). Amsterdam: John Benjamins.

Fougeron, C., Smith, C.L. (1993). Illustrations of the IPA: French, *Journal of the International Phonetic Association* 23: 73-76.

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological cybernetics*, *72*(1), 43-53

Hualde, J. I. (2005). *The sounds of Spanish*. Cambridge University Press.

Kamiyama, T., Vaissière, J. (2009). Perception and production of French close and close-mid rounded vowels by Japanese-speaking learners. *Acquisition et interaction en langue étrangère* Aile... Lia 2, 9-41.

Kamiyama, T., Detey, S., Kawaguchi, Y. (2017). Les japonophones. In S. Detey, I. Racine, Y. Kawaguchi and J. Eychenne (Eds.), *La prononciation du français dans le monde : du natif à l'apprenant,* CLE International, pp. 155-161.

Kocjančič Antolík, T., Pillot-Loiseau, C., Kamiyama, T. (2019). The effectiveness of real-time ultrasound visual feedback on tongue movements in L2 pronunciation training: Japanese learners improving the French vowel contrast /y/-/u/, *Journal of Second Language Pronunciation, 5*(1), 72-97.

Kröger, B. J., & Kannampuzha, J. (2008). A neurofunctional model of speech production including aspects of auditory and audio-visual speech perception. In *Auditory Visual Speech Processing Conference* AVSP (pp. 83-88).

Kühnert, B., & Pillot-Loiseau, C. (2022). Teaching Pronunciation with Direct Visual Articulatory Feedback: Pedagogical Considerations for the Use of Ultrasound in the Classroom, *RANAM (Recherches Anglaises et Nord-américaines), 55,* 9-24.

Li, J. J., Ayala, S., Harel, D., Shiller, D. M., & McAllister, T. (2019). Individual predictors of response to biofeedback training for second-language production. *The Journal of the Acoustical Society of America, 146*(6), 4625-4643.

Meadows, B. (2007). Implications of ultrasound technology in the L2 classroom. *Journal of Second Language Acquisition and Teaching*, *14*, 15-41.

Racine, I. (2017). Les hispanophones. In S. Detey, I. Racine, Y. Kawaguchi and J. Eychenne (Eds.), *La prononciation du français dans le monde : du natif à l'apprenant,* CLE International, pp. 143-148.

Racine, I., & Detey, S. (2018). Production of French close rounded vowels by Spanish learners A corpus-based study. In Mark Gibson and Juana Gil (Eds), *Romance phonetics and phonology*, Oxford: Oxford University Press, pp. 381-394.

Vance, T. J. (2008). *The sounds of Japanese*. Cambridge: Cambridge University Press.

Pinocchio, a biomimetic mechatronic testbed for producing *in vitro* articulated speech

Nathalie Henrich Bernardoni[1], Julien Royer[1], Mounib Tlaidi[1],
Xavier Laval[1], Sylvain Arnaud[1], Lucie Bailly[2]

[1]*Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France*
[2]*Univ. Grenoble Alpes, CNRS, Grenoble INP, 3SR, 38000 Grenoble, France*
nathalie.henrich@gipsa-lab.fr, lucie.bailly@3sr-grenoble.fr

Introduction.
Speech production results from a fine neuro-controlled coordination between breathing, phonatory and articulatory movements. In physical terms, these movements induce fluid-structure-acoustic interactions within the vocal tract. In the framework of source-filter theory (Fant 1960), the aerodynamic phase of speech is often overlooked, yet being of much importance for voiced and unvoiced speech sound production (Catford 1977). In the case of voiced speech sound, it is taken into account by myoelastic-aerodynamic theory of phonation (Jan G. Švec et al. 2021), in which glottal constriction and vocal folds vibration generate aeroacoustic sources. Our aim is to advance our understanding of speech sound production by developing a biomimetic *in vitro* testbed. Over the last thirty years, vocal-fold testbeds have evolved in complexity and biomimicry. However, most testbeds explore the physics of phonation on geometrically-fixed replicas capable of self-sustained oscillations in fluid-structure interaction, but unable to produce intonative variations. Most of the time, the testbeds are not coupled to vocal tract, and whenever they are, the resonant cavities are static 3D-printed tracts. Reproducing *in vitro* the dynamic movement of speech articulators, such as jaw, tongue, velum and larynx, together with phonation, remains a challenge. We present here the first steps in the design of a biomimetic mechatronic testbed that would integrate all phonatory and articulatory aspects important for voice and speech production.

Methods.
Design efforts focused on three aspects : how to breath, how to phonate, how to articulate.
The first aspect to take into account is the ability to control air supply through flow and pressure parameters. The system controls a valve opening, which is linearly related to airflow rate at the device inlet. Subglottal pressure can also be programmed using a servo loop. To stabilise the upstream jet, a settling chamber is used. A long tube leading from the chamber to the testbed avoids acoustic coupling with the subglottal tract (Lehoux, Hampala, and Jan G Švec 2021).
Choice was made to design a flexible laryngeal envelope combined with a 1:1 scale vocal-folds replica, so as to allow articulatory movements within the larynx and vocal fold stretching. In a first approach, the folds were designed to be inserted interchangeably while maintaining a seal, with the aim of gradually improving their biomimetic characteristics (Luizard et al. 2023). In a second approach, the folds were moulded together with the laryngeal envelope. Several aspects were taken into account and tailored : folds geometry, material tensile stiffness, folds anisotropy. Folds geometry was derived from common M5 models (Murray and Thomson 2012; Luizard et al. 2023). All tested folds were made of either commonly-used silicone elastomers (Ecoflex™series with increasing degrees of shore hardness) or gelatin-based hydrogels. Body and cover were homogenous (moulded together) or heterogenous (moulded apart with different stiffness, and without or with addition of a fibrous structure). The self-oscillation capabilities of each laryngeal replica were assessed.
First implemented articulatory movements were within the larynx itself, in the framework of Laryngeal Articulator Model (Esling et al. 2019). The vocal folds can be adducted or abducted, imitating the phases of breathing and phonation. They can be stretched and compressed into the anterior-posterior direction, and compressed or uncompressed into the medial-lateral one. Secondly, a geometrically-realistic and articulatory-driven vocal tract was designed, inspired from Arai's vocal-tract models (Arai 2016). A FE model of the tongue was used to design the tongue mould (Hermant, Perrier, and Payan 2017). Two primary articulatory movements were implemented : mandibular motion and shaping of the oral cavity with the movement of tongue body and apex.

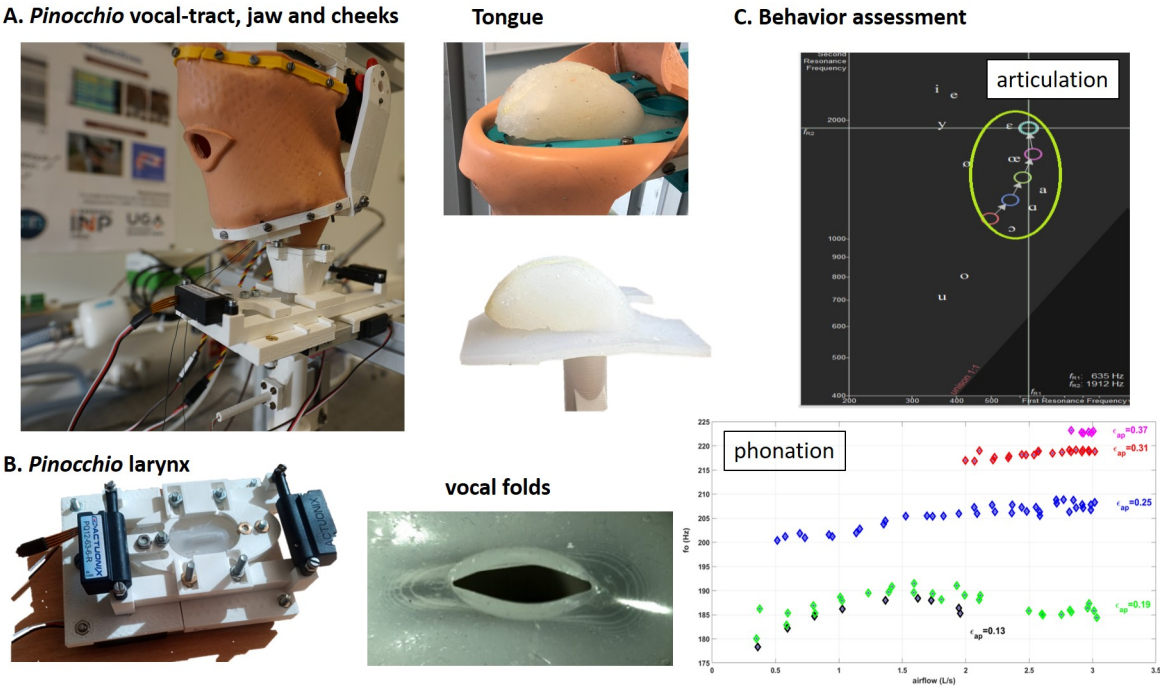


Figure 1: *Presentation of the pinocchio testbed.*

Results and discussion.
Figure 1 presents the the complete system, called *pinocchio*,and its phonatory and articulatory capabilities.
Phonation The vocal-folds replica were able to self-oscillate with and without a vocal-tract, over a wide range of flow rates and for different degrees of tensile stretch. First main factor of variability in vibratory behaviour was the folds geometry. While long ($20mm$ at rest) and thick ($4mm$) folds did not allow a self-oscillation frequency in the expected range of human speech (Luizard et al. 2023), shortening the length at rest (down to maximun $12mm$) and slimming the fold (down to $1.5mm$) achieved the goal of a self-oscillation frequency up to the range of female speech (f_o from $180to220Hz$) as shown in Figure 1C. The latest version of the vocal-fold replica allowed for increasing f_o with increasing pre-strain in the anterior-posterior direction (ϵ_{ap}).
Articulation We explored the vocalic space produced by movements of jaw, tongue body and apex movements (see Figure 1C top panel). F1 varied moderately between 500 and $700Hz$, mainly in relation to mouth opening induced by jaw movement. F2 varied between 1000 and $2000Hz$ with tongue and jaw movements. At this early stage of his life, Pinocchio was able to produce vowels in the surrounding of [a].

References.

Arai, Takayuki (2016). "Vocal-tract models and their applications in education for intuitive understanding of speech production". In: *Acoustical Science and Technology* 37.4, pp. 148–156.

Catford, John Cunnison (1977). *Fundamental Problems in Phonetics*. Indiana University Press.

Esling, John H, Scott R Moisik, Allison Benner, and Lise Crevier-Buchman (2019). *Voice quality: The laryngeal articulator model*. Vol. 162. Cambridge University Press.

Fant, G. (1960). *Acoustic Theory Of Speech Production*. The Hague: Mouton.

Hermant, Nicolas, Pascal Perrier, and Yohan Payan (2017). "Human tongue biomechanical modeling". In: *Biomechanics of Living Organs*. Elsevier, pp. 395–411.

Lehoux, Hugo, Vít Hampala, and Jan G Švec (2021). "Subglottal pressure oscillations in anechoic and resonant conditions and their influence on excised larynx phonations". In: *Scientific reports* 11.1, pp. 1–14.

Luizard, P, L Bailly, H Yousefi-Mashouf, R Girault, L Orgéas, and N Henrich Bernardoni (2023). "Flow-induced oscillations of vocal-fold replicas with tuned extensibility and material properties". In: *Scientific reports* 13.22658, pp. 1–19.

Murray, Preston R. and Scott L. Thomson (2012). "Vibratory responses of synthetic, self-oscillating vocal fold models". In: *The Journal of the Acoustical Society of America* 132.5, pp. 3428–3438. (Visited on 09/21/2022).

Švec, Jan G., Harm K. Schutte, C. Julian Chen, and Ingo R. Titze (2021). "Integrative Insights into the Myoelastic-Aerodynamic Theory and Acoustics of Phonation. Scientific Tribute to Donald G. Miller". In: *Journal of Voice*.

A Biomechanical Tongue Model of a Neanderthal

Maxime CALKA[1], Pablo ALVAREZ[1], Pascal PERRIER[2], Yohan PAYAN[3], Amélie VIALET[1,4]

[1] Sorbonne Université, Institut des Sciences du Calcul et des Données, Paris, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
[3] Univ. Grenoble Alpes, CNRS, Grenoble INP, TIMC, 38000 Grenoble, France
[4] Muséum national d'Histoire naturelle, UMR 7194 - Histoire naturelle de l'Homme préhistorique, Paris, France

maxime.calka@sorbonne-universite.fr, pascal.perrier@gipsa-lab.grenoble-inp.fr, yohan.payan@univ-grenoble-alpes.fr, amelie.vialet@mnhn.fr

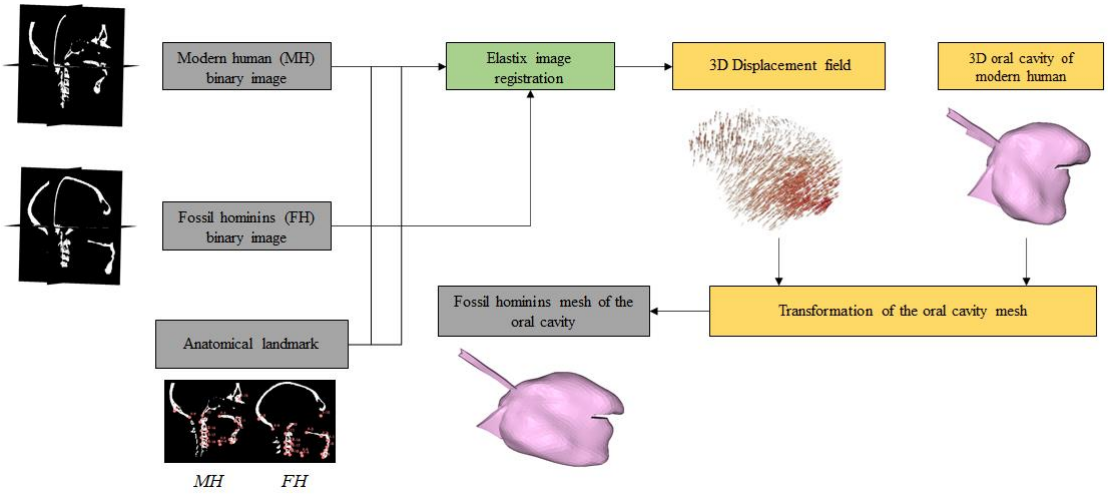
Introduction. The emergence of the capacity for spoken language in humans during the course of human evolution is a widely debated question. The complexity of this question lies in the difficulty of studying fossil hominins due to the poor preservation of the phonatory apparatus. Soft tissues and cartilage do not fossilize, while bones can be damaged, deformed, or eroded. Qualitative observations currently support the idea that fossil hominins had the ability to speak (Steele et al., 2012; d'Errico & Colagè, 2022). However, no quantitative measurements have been able to provide conclusive evidence in this direction. In this scientific context, our long-term project aims to provide a quantitative evaluation of the suitability of the biological characteristics of fossil hominins for articulated speech. Our approach consists of three main steps: (1) predicting the morphology of the missing tongue and surrounding soft tissues in the oropharyngeal cavity from the geometry of the skull, mandible, and vertebrae; (2) constructing a biomechanical model of the predicted fossil tongue and its surrounding structures in the oral cavity, incorporating the muscles responsible for its movements and shaping; (3) using this model to evaluate the maximal movement magnitudes of the tongue in the anterio-posterior and vertical dimensions, the range of variation of achievable vocal tract shapes, and the capacity of fossil hominins to maintain stable, differentiated tongue postures, providing a basis for the production of distinctive articulated sounds. Because it incorporates a realistic representation of the intrinsic physical characteristics of the tongue and its neurophysiological control, such as muscle anatomy and control, this work allows us to surpass previous modeling studies that have relied solely on geometrical reconstructions and models of the vocal tract of fossil hominins (Boë et al., 2002; Boë et al., 2007) to offer initial evidence supporting their ability to produce distinct vowels.

Methods. To generate the biomechanical model of fossil hominins, we utilized a method designed for the automatic generation of finite-element biomechanical models of both human and non-human primates. This method involves morphing a reference model of a living male human subject. It has been carefully evaluated for its ability to generate an accurate model of a Baboon tongue, chosen for evaluation due to its significant morphological differences from humans (Alvarez et al., 2024). The method relies on a 3D binary-image registration technique found in the Elastix library (Klein et al., 2009). This technique utilizes 3D X-Ray scans of the head and neck region and accomplishes two successive major morphological registrations: (1) an affine registration aligning the two skulls, and (2) a non-rigid "B-Spline" registration offering more detailed transformations within the skull (Bijar et al., 2016). The displacement field generated by this two-step registration process is then applied to the finite element model of the tongue of a living human. Further details regarding this model, including its topology, mesh, materials, muscular model, etc., can be found in Calka (2023). The process of creating this tongue model is illustrated in Figure 1 (upper panel). In the current study, it has been applied to a homo neanderthalensis known as "La Ferrassie 1" or LF1 (male, 70-50 ka).

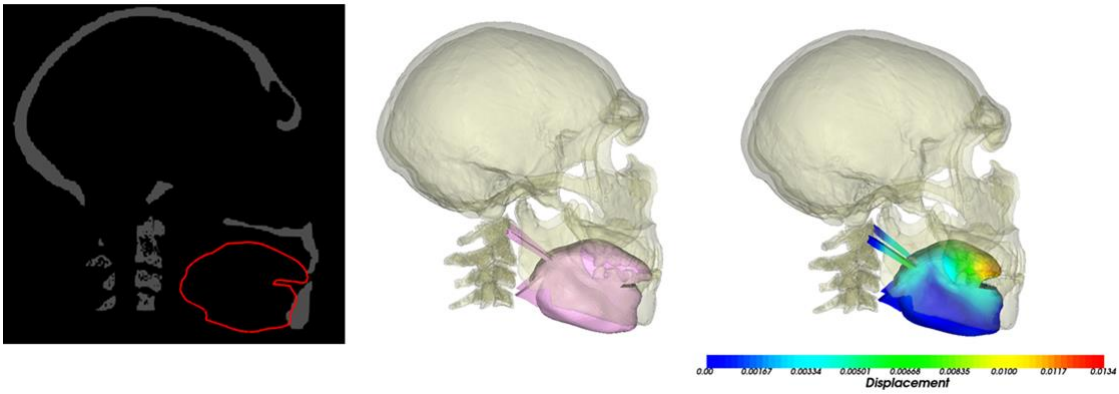
Results. The obtained tongue model for LF1 is depicted in Figure 1 (Panel B, left). Additionally, the lower panel of Figure 1 displays the 3D model of the Neanderthal skull (middle) alongside the outcome of a simulation where the tongue muscles are activated (right) as they would be during the production of the vowel /u/. Compared with the tongue of LF1, the tongue of this sapiens (Panel A) is flatter and more elongated, rather like that of a non-human primate.

Discussion. It is obviously impossible to quantitatively validate the accuracy of the model. However, the validation of the method provided in Alvarez et al. (2024) encourages us to have confidence in the predictions. Moreover, the overall shape of the predicted tongue within the oral cavity (Figure 1, lower panel) appears plausible, aligning with observations from non-human primates (Riede et al., 2005). In terms of shapes and dimensions, our results appear to be consistent with those reported by Boë et al. (2013). The strength of our biomechanical modeling approach lies in its ability to facilitate the examination of tongue posture stability amidst variations in muscle activations. Additionally, it allows for exploring the sensitivity of the range of articulated sounds to different mechanical parameters, such as the stiffness of tongue tissues, the muscle force generation capability, or the position of the hyoid bone. In the short term, simulating the production of

consonants and critical vowels, such as /i/, which appears absent in non-human primates, will offer fresh insights into the development of speech in Neanderthals.



Panel A



Panel B

Figure 1: *Panel A - The process of creating a fossil hominin tongue model. Panel B - La Ferrassie 1's tongue model. Left: Generated tongue (in red) superimposed on volumic image of his skull. Middle: Mesh of the La Ferrassie 1's tongue inside his skull. Right: Simulation of a quasi-/u/ phoneme using La Ferrassie 1's.*

References

Alvarez, P., El Mouss, M., Calka, M., Belme, A., Berillon, G., Brige, P., & Vialet, A. (2024). Predicting primate tongue morphology based on geometrical skull matching. A first step towards an application on fossil hominins. PLOS Computational Biology, 20(1), e1011808.

Barney, A., Martelli, S., Serrurier, A., & Steele, J. (2012). Articulatory capacity of Neanderthals, a very recent and human-like fossil hominin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1585), 88-102.

Bijar, A., Rohan, P. Y., Perrier, P., & Payan, Y. (2016). Atlas-based automatic generation of subject-specific finite element tongue meshes. Annals of biomedical engineering, 44, 16-34.

Boë, L. J., Heim, J. L., Honda, K., & Maeda, S. (2002). The potential Neandertal vowel space was as large as that of modern humans. *Journal of Phonetics*, *30*(3), 465-484.

Boë, L. J., Heim, J. L., Honda, K., Maeda, S., Badin, P., & Abry, C. (2007). The vocal tract of newborn humans and Neanderthals: Acoustic capabilities and consequences for the debate on the origin of language. A reply to Lieberman (2007a). *Journal of Phonetics*, *35*(4), 564-581.

Calka, M. (2023). Modélisation biomécanique par éléments finis de la langue : évaluation, production de la parole et perspectives d'application à la chirurgie linguale assistée par ordinateur (Doctoral dissertation, Université Grenoble-Alpes).

d'Errico, F., & Colagè, I. (2022). The emergence of symbolic cognition. In *The Routledge International Handbook of Neuroaesthetics* (pp. 539-554). Routledge.

Klein, S., Staring, M., Murphy, K., Viergever, M. A., & Pluim, J. P. (2009). Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, *29*(1), 196-205.

Lieberman, P., Crelin, E. S., & Klatt, D. H. (1972). Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. *American Anthropologist*, *74*(3), 287-307.

Riede, T., Bronson, E., Hatzikirou, H., & Zuberbühler, K. (2005). Vocal production mechanisms in a non-human primate: morphological data and a model. *Journal of Human Evolution*, *48*(1), 85-96.

Steele, J., Ferrari, P. F., & Fogassi, L. (2012). From action to language: comparative perspectives on primate tool use, gesture and the evolution of human language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1585), 4-9.

Acquisition of articulatory dynamics in second language speech: Japanese speakers' production of English and Japanese liquids

Takayuki Nagamine[1]

[1]*Department of Linguistics and English Language, Lancaster University, UK.*
t.nagamine@lancaseter.ac.uk

Introduction. First-language (L1) Japanese speakers have difficulty in producing English liquids. It is widely hypothesised that this is a consequence of L1 Japanese speakers classifying English liquids as poor instances of the L1 liquid category (Japanese /ɾ/; Bradlow 2008). The precise nature of this influence in articulation, however, is not widely understood, due to a lack of direct comprehensive research on English versus Japanese liquids. Given that alveolar taps, the canonical realisation of Japanese /ɾ/, show a greater vocalic coarticulation than other members of liquids, I hypothesise that (1) L1 Japanese speakers show distinct liquid-vowel coarticulatory patterns in English liquids according to vowel contexts due to a carry-over effect from articulatory strategies for Japanese /ɾ/ and (2) proficient L2 learners are more capable of adjusting coarticulation patterns than less proficient L2 learners (Beristain 2022).

Methods. Midsagittal ultrasound tongue images and audio recordings were collected from 29 L1 Japanese speakers and 14 L1 North American English speakers. L1 Japanese speakers are classified into 'intermediate' (n = 9) and 'advanced' (n = 20) groups based on perceptual identification accuracy of English /l ɹ/ using Gaussian mixture models. Target words for the production study include 16 English words (eight minimal pairs) contrasted by word-initial /l/ and /ɹ/ and five Japanese words with word-initial Japanese /ɾ/ preceding /a/, /i/ and /u/. In total, there are 1,309 tokens of English /l/, 1,321 tokens of English /ɹ/ and 445 tokens of Japanese /ɾ/ for analysis. Tongue splines were tracked based on a set of x/y Cartesian coordinates for 11 reference points along the tongue surface using the DeepLabCut plug-in via Articulate Assistant Advanced (AAA), which were then within-speaker z-scored for cross-speaker comparison. Tongue splines were extracted in the analysis window consisting of acoustically delimited word-initial liquid-vowel intervals with an additional 350ms interval padded before the liquid onset (as articulatory onset can precede acoustic onset).
Three statistical analyses were used to evaluate liquid-vowel coarticulatory patterns. First, I run Principal Component Analysis (PCA) to identify key dimensions in midsagittal tongue shape throughout the interval and summarise them into numeric values (PC scores). I then conduct *functional* PCA (FPCA) to further convert into numeric values (FPC scores) time-varying changes of the PC scores from 350ms prior to the onset through to the offset of the word-initial liquid-vowel interval. Finally, I fit Bayesian hierarchical regression models to predict the PC trajectory patterns, expressed numerically by FPC scores, by vowel contexts and groups as fixed effects and the interaction between them. The random effects include by-speaker varying intercepts and slopes for vowel contexts and by-item varying intercepts and slopes for groups. I use weakly informed priors to allow for a wide range of possible values. I report the estimated coefficients (β) and the 95% credible intervals (i.e., [*min.*, *max.*]) calculated from the model (cf. Roettger, Mahrt, and Cole 2019).

Results. The PCA analysis is shown in the left panel in Figure 1. The first two PCs explain the largest variation in the data: 39.28% for PC1 and 30.59% for PC2. PC1 is associated with tongue body raising and lowering and PC2 with tongue advancement. In the subsequent analyses, I focus on PC1 given its proportion of variance being the largest.
Next, the results of the FPCA analysis is shown in the middle panel in Figure 1 for English /l/ (top), English /ɹ/ (middle) and Japanese /ɾ/ (bottom). Here, to highlight the overall tendency, time-varying changes of PC1 scores are reconstructed directly from FPC1 that accounts for the largest variation in trajectory pattern (58.10%). The x-axis represents proportional time from the onset of the 350ms window (0%) to the vowel offset (100%), and the two vertical dotted lines represent mean liquid interval. The trajectories show that the speaker's tongue position is close to the mean tongue shape prior to the liquid onset, representing their pre-speech posture (Wilson and Kanada 2014). The tongue is then slightly retracted immediately before transitioning into the liquid and vowel portion. During the liquid and vowel interval, the trajectory for the /i/ context (red) is associated with higher FPC1 values, indicating a greater degree of tongue body raising and fronting

(i.e., higher PC1 values), followed by in the /u/ (green) and /a/ (blue) contexts. Here, a clear clustering of dynamic PC1 contours is shown for L1 Japanese speakers (i.e., intermediate and advanced groups), suggesting that tongue body movement across vowel contexts is more variable for L1 Japanese speakers than for L1 English speakers.

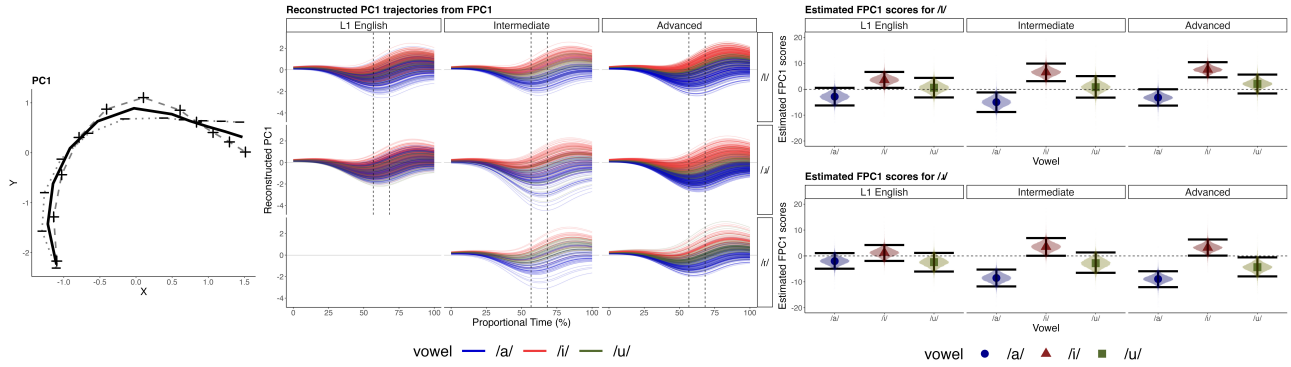


Figure 1: *Left: Variation in midsagittal tongue shape captured by PC1. Middle: Time-varying PC1 changes reconstructed by FPC1. Right: Posterior distributions of FPC1 values for English /l/ (top) and /ɹ/ (bottom).*

Finally, the FPC1 values estimated from the Bayesian hierarchical regression models are shown in the right panel in Figure 1. FPC1 values are overall higher in the /i/ context, followed by the /u/ and /a/ contexts. English /ɹ/ shows a group-vowel interaction for the estimated FPC1 scores. Credible intervals for L1 English speakers show greater overlap with zero (β = 1.26 [−1.93, 4.22] for /i/, β = −1.99 [−4.93, 1.10] for /a/ and β = −2.40 [−6.03, 1.17] for /u/). This indicates that their productions are more similar across vowel contexts compared to that of L1 Japanese speakers (β = 3.55 [0.07, 6.91] for /i/, β = −8.51 [−11.78, 6.91] for /a/ and β = −2.79 [−6.51, 1.35] for /u/ for the intermediate group; β = 3.22 [0.13, 6.36] for /i/, β = −8.92 [−12.05, −5.91] for /a/ and β = −4.34 [−7.90, −0.55] for /u/ for the advanced group).
English /l/, on the other hand, does not show clear group-vowel interactions. Credible intervals show a greater degree of overlap with zero for /u/ (β = 0.59 [−3.17, 4.38] for L1 English speakers, β = 0.79 [−3.24, 5.07] for intermediate L1 Japanese speakers, and β = 2.05 [−1.63, 5.66] for advanced L1 Japanese speakers) compared to /i/ (β = 3.59 [0.53, 6.68] for L1 English speakers, β = 6.57 [3.12, 9.91] for intermediate L1 Japanese speakers and β = 7.59 [4.60, 10.45] for advanced L1 Japanese speakers) or /a/ (β = −2.87 [−6.27, 0.53] for L1 English speakers, β = −4.99 [−8.79, −1.23] for intermediate L1 Japanese speakers and β = −3.22 [−6.32, −0.02] for advanced L1 Japanese speakers).

Discussion and Conclusion. The analysis demonstrates that L1 Japanese speakers show a greater liquid-vowel coarticulation than L1 English speakers for English /ɹ/, suggesting L1 transfer of liquid-vowel coarticulation into L2 (cf. Yamane, Howson, and Po-Chun (Grace) 2015). Vowel contexts influence the production of English /l/ in a similar manner across groups, which may reflect a weaker coarticulatory resistance for English /l/ compared to English /ɹ/ (Proctor et al. 2019). No notable effects of L2 proficiency, investigated through classifying L1 Japanese speakers into 'intermediate' and 'advanced' groups using perceptual accuracy, are found. This could be because perceptual accuracy does not well capture differences in coarticulatory adjustability among L1 Japanese speakers who are relatively homogeneous late bilinguals (Beristain 2022). Future research will explore the articulatory dimensions not included in this study.

References.

Beristain, Ander Murillo (2022). "The Acquisition of Acoustic and Aerodynamic Patterns of Coarticulation in Second and Heritage Languages". PhD thesis. University of Illinois Urbana-Champaign.

Bradlow, Ann R (2008). "Training Non-Native Language Sound Patterns". In: *Phonology and Second Language Acquisition*. Ed. by Jette G. Hansen Edwards and Mary Zampini L. John Benjamins Publishing Company, pp. 287–308.

Proctor, Michael, Rachel Walker, Caitlin Smith, Tünde Szalay, Louis Goldstein, and Shrikanth Narayanan (2019). "Articulatory Characterization of English Liquid-Final Rimes". In: *Journal of Phonetics* 77, p. 100921. DOI: 10.1016/j.wocn.2019.100921.

Roettger, Timo B., Tim Mahrt, and Jennifer Cole (2019). "Mapping Prosody onto Meaning – the Case of Information Structure in American English". In: *Language, Cognition and Neuroscience* 34.7, pp. 841–860. DOI: 10.1080/23273798.2019.1587482.

Wilson, Ian and Sunao Kanada (2014). "Pre-Speech Postures of Second-Language versus First-Language Speakers". In: *Journal of the Phonetic Society of Japan* 18.2, pp. 106–109. DOI: 10.24467/onseikenkyu.18.2_106.

Yamane, Noriko, Phil Howson, and Wei Po-Chun (Grace) (2015). "An Ultrasound Examination of Taps in Japanese". In: *Proceedings of the 18th International Congress of Phonetic Sciences*. Ed. by The Scottish Consortium for ICPhS 2015. Glasgow, UK: The International Phonetic Association, pp. 1–5.

A dynamic neural field model of vowel diphthongisation

Sam Kirkham — Lancaster University — s.kirkham@lancaster.ac.uk

Patrycja Strycharczuk — University of Manchester — p.strycharczuk@manchester.ac.uk

Introduction.
The variable diphthongisation of vowels in English is a widely attested form of synchronic variation, such as the monothongisation of GOAT and PRICE in the dialects of Northern England (Hughes, Trudgill, and Watt 2012), as well as diphthongisation of tense monophthongs, such as FLEECE and GOOSE (Strycharczuk, López-Ibáñez, et al. 2020; Wells 1982). Diphthongisation also underlies many diachronic sound changes, such as the development of high vowels into diphthongs during the English Great Vowel Shift (Jespersen 1909), which appears to involve splitting a single long vowel into a two-target diphthong. While these descriptive facts of vowel variation are well-documented, it remains challenging to provide a convincing account of vowel diphthongisation that can capture the wide range of gradient synchronic variation in dialects and the apparently categorical shifts of long-term sound change. In this paper, we develop a theoretical account of vowel variation and change, grounded in a dynamic neural field account of speech planning (Tilsen 2016; Roon and Gafos 2016) that feeds into a task dynamic model of articulatory execution (Saltzman and Munhall 1989).

Methods.
First, we outline a model that proposes a compositonal two-target structure for all long vowels, including long monophthongs and diphthongs (Strycharczuk, Kirkham, et al. submitted). In this view, a long monophthong is long because it is comprised of two sequentially-timed gestures, each of which has identical targets. A diphthong has the same underlying structure (two targets), but has different target parameters for each of the targets, thus yielding movement from the first target to the second. We illustrate this using task dynamic simulations based on the model in Sorensen and Gafos (2016), specifying a vowel as two concatenated gestural activation intervals of 250 ms in duration, which are coupled anti-phase to one another. Our model predicts that variation between a long monophthong and a diphthong can be captured entirely via gradient variation in the nucleus target value.

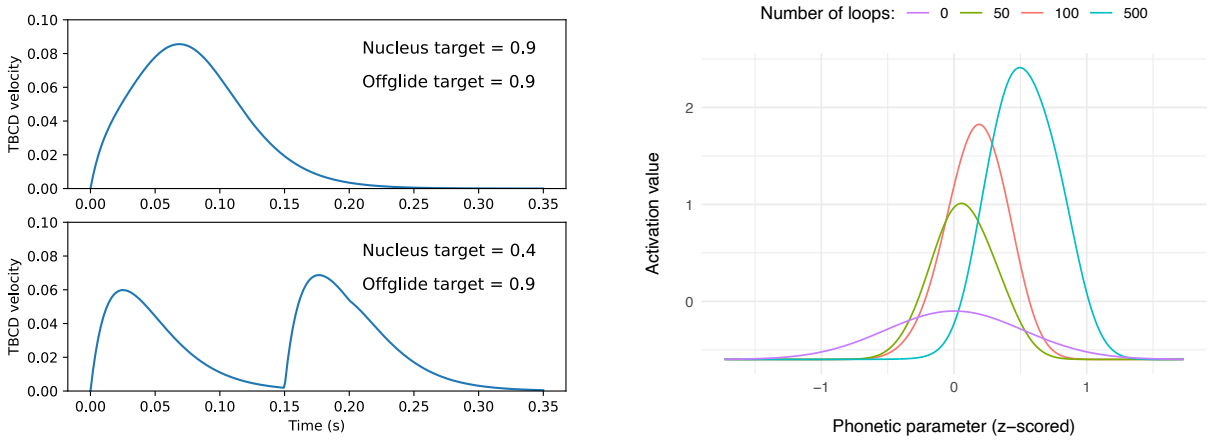


Figure 1: *LEFT: Velocity of simulated gestures with two identical targets (top) and two different targets (bottom). RIGHT: Example DFT production-perception simulation.*

Our second analysis focuses on how the phonological representations of individual speakers change during a sound

change. We advance a dynamic neural field (DNF) model (Schöner, Spencer, and The DFT Research Group 2016), which has proven a versatile tool for dynamical models of phonological planning (Kirov and Gafos 2007; Roon and Gafos 2016; Tilsen 2019; Shaw and Tang 2023). A DNF model situates phonological planning in an activation field over a range of phonetic parameters. A dynamical equation specifies the evolution of field activation until some value reaches a threshold, which is then selected as the parameter value for speech production. We then model production and perception as inputs to the field and track how the field develops over time.

Results.
Figure 1 demonstrates how one versus two velocity minima can result from variation in the simulated nucleus target, despite both simulations containing the same gestural activation intervals timed in exactly the same way. We show that such simulations also generate a gradient continuum between a monophthong and a diphthong, providing a clear mechanism for variation and change. Following on from this, our DFT model then defines /i/ as two planning fields (one for the nucleus, one for the offglide), which are specified as a pair of coupled differential equations with inhibitory and excitatory components. We model production-perception as (i) a speaker producing a value from their activation field; (ii) hearing a speaker whose nucleus has a phonetic bias towards /e/; (iii) this perceived token is integrated into memory with a small amount of noise; (iv) this process repeats (Kirov and Gafos 2007). After a number of interactions with this 'biased' speaker, the activation field shifts away from the initial state (representing an /i/ nucleus) towards a new peak (representing an /e/-like nucleus), as in Figure 1. We simulate articulatory trajectories based on these activation fields and show that /i/ eventually changes into /ai/, with no recourse to categorical rules. Specifically, when the vowel nucleus moves away from its initial state towards a different state, the outcome of the task dynamic equation is more likely to be a diphthong.

Discussion.
Our model combines an autonomous model of gestural dynamics with a dynamic field planning model in order to simulate the processes of diachronic vowel diphthongisation. We propose that the accumulation of gradient variation in targets can lead to long-term sound changes that look like categorical changes over historical time. In conclusion, we identify a shared mechanism for synchronic variation and diachronic change in vowels – gradient variation in gestural targets – and propose a mechanism for how individual phonological representations change. We also discuss potential variability in both vowel targets as part of a broader stochastic model of sound change.

References.

Hughes, Arthur, Peter Trudgill, and Dominic Watt, eds. (2012). *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles*. Fifth. London: Hodder.

Jespersen, Otto (1909). *A Modern English Grammar on Historical Principles*. London: George Allen & Unwin Ltd.

Kirov, Christo and Adamantios I. Gafos (2007). "Dynamic phonetic detail in lexical representations". In: *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 637–640.

Roon, Kevin D. and Adamantios I. Gafos (2016). "Perceiving while producing: Modeling the dynamics of phonological planning". In: *Journal of Memory and Language* 89.2, pp. 222–243.

Saltzman, Elliot and Kevin G. Munhall (1989). "A dynamical approach to gestural patterning in speech production". In: *Ecological Psychology* 1.4, pp. 333–382.

Schöner, Gregor, John P. Spencer, and The DFT Research Group (2016). *Dynamic Thinking: A Primer on Dynamic Field Theory*. Oxford: Oxford University Press.

Shaw, Jason A. and Kevin Tang (2023). "A dynamic neural field model of leaky prosody: proof of concept". In: *Proceedings of the Annual Meeting on Phonology 2022*, pp. 1–12.

Sorensen, Tanner and Adamantios I. Gafos (2016). "The gesture as an autonomous nonlinear dynamical system". In: *Ecological Psychology* 28.4, pp. 188–215.

Strycharczuk, Patrycja, Sam Kirkham, Emily Gorman, and Takayuki Nagamine (submitted). "Towards a dynamical model of English vowels: Evidence from diphthongisation". In.

Strycharczuk, Patrycja, Manuel López-Ibáñez, Georgina Brown, and Adrian Leemann (2020). "General Northern English: Exploring regional variation in the north of England with machine learning". In: *Frontiers in Artificial Intelligence* 3.48, pp. 1–18.

Tilsen, Sam (2016). "Selection and coordination: The articulatory basis for the emergence of phonological structure". In: *Journal of Phonetics* 55, pp. 53–77.

— (2019). "Motoric mechanisms for the emergence of non-local phonological patterns". In: *Frontiers in Psychology* 10.2143, pp. 1–25.

Wells, John C. (1982). *Accents of English: Volumes 1–3*. Cambridge: Cambridge University Press.

An experimental setup for capturing multimodal accommodation using dual electromagnetic articulography, audio, and video

Lena Pagel[1], *Simon Roessig*[2], *Doris Mücke*[1]

[1]University of Cologne, Germany
[2]University of York, United Kingdom
lena.pagel@uni-koeln.de, simon.roessig@york.ac.uk, doris.muecke@uni-koeln.de

Introduction. Previous research has shown that speakers frequently accommodate to their interlocutor's speech patterns and speech-accompanying movements, e.g. in facial expression (Louwerse *et al.* 2012), manual gestures (Mol *et al.* 2012), intonation (Babel & Bulatov 2012), speaking rate (Cummins 2002), and acoustic properties of segment productions (Nielsen 2011). Facilitated by recent technological advances, a small number of studies have used dual electromagnetic articulography (dual EMA) to provide a clearer understanding of accommodation in supra-laryngeal articulation (Lee *et al.* 2018; Mukherjee *et al.* 2018; Tiede & Mooshammer 2013). However, to date, not many studies have integrated the multiple modalities of accommodation within one experimental setting (but cf. Duran & Fusaroli 2017; Louwerse *et al.* 2012; Oben & Brône 2016). Moreover, to our knowledge, only one study has analysed multimodal accommodation using dual EMA, but it included only one dyad of speakers (Tiede *et al.* 2010). Furthermore, elicitation methods are highly diverse across previous studies and frequently do not take information structure of utterances into account. Information structure, however, affects the production of speech and co-speech motion within a speaker and can also influence the amount of accommodation between speakers (Lee *et al.* 2018), underlining its relevance in experimental designs. Here, we present a methodological approach to capture multimodal accommodation using a setup with dual EMA, audio, and video, which can inform future studies concerning structure, task, and technical setup. We introduce the cooperative game *DiCE* to elicit lexically and prosodically controlled data in an engaging setting, which is available at https://osf.io/9fmqh/.

Methods. Speakers are recorded in dyads, intentionally paired with unfamiliar partners. For each dyad, the recording session includes a solo condition for each speaker individually and a dialogue condition for both speakers together (see Figure 1 for photos). The card game *DiCE* (*Dialogic Collecting Expedition*) is designed to elicit the production of lexically and prosodically controlled utterances. Participants work together to collect and sort valuable items from across the world. They interact through question-answer sets with a fixed lexical structure, aiming to discover which cards they are holding. One participant's question elicits the focus structure of the other participant's answer. The speech material includes target words for objects (*Bohne, Mode, Vase, Made*; Engl. *bean, fashion, vase, maggot*) and cities (*Manila, Medina, Benali, Milano*), which occur either in corrective focus or in the background. Additionally, speakers produce pointing gestures to indicate the location of the intended card. In the solo condition preceding the dialogue condition, each speaker plays a simplified digital version of the game by answering questions prompted on a screen, with the other speaker absent from the room. The participants are recorded with dual 3D EMA (one articulograph per speaker, each with 16 sensors attached for capturing speech and co-speech kinematics), head-mounted microphones (one per speaker) and three cameras (one per speaker from the front plus one from the side). This setup allows analyses of (i) acoustic speech cues (audio signal: F0, intensity, spectral properties of consonant and vowel productions), (ii) vocal tract kinematics (EMA signal: lip aperture, lip spreading, jaw, tongue tip, and tongue body movements), (iii) kinematics of speech-accompanying body movements (EMA signal: head motion, eyebrow raising and furrowing, torso and shoulder movement; video signal: facial expression, pointing gestures), and (iv) kinematics of smiles and breathing (EMA signal: lip spreading, torso expansion; video signal: smiles). We have successfully implemented the methodological approach in recordings of 15 German-speaking dyads, to our knowledge forming the largest existing corpus of dual EMA recordings.

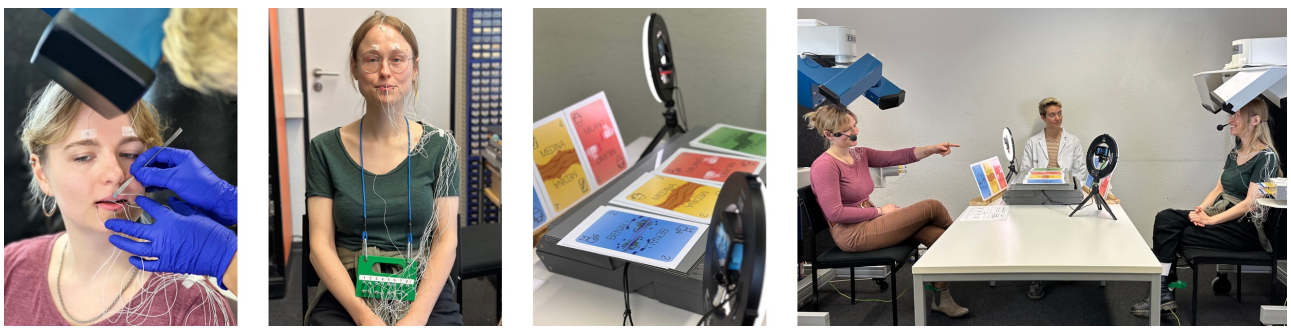


Figure 1: *From left to right: EMA sensor placement on one speaker during preparation; the other speaker with portable sensor mounting during preparation; DiCE game during dialogue; recording setup during dialogue.*

Figure 2 exemplarily shows the recorded multimodal data for one question-answer set in the dialogue condition of one dyad. Four parameters of speech and co-speech kinematics (lip aperture, vertical tongue, head, and eyebrow motion) are selected from the wide range of possible parameters to exemplify the nature of the recorded multimodal data.

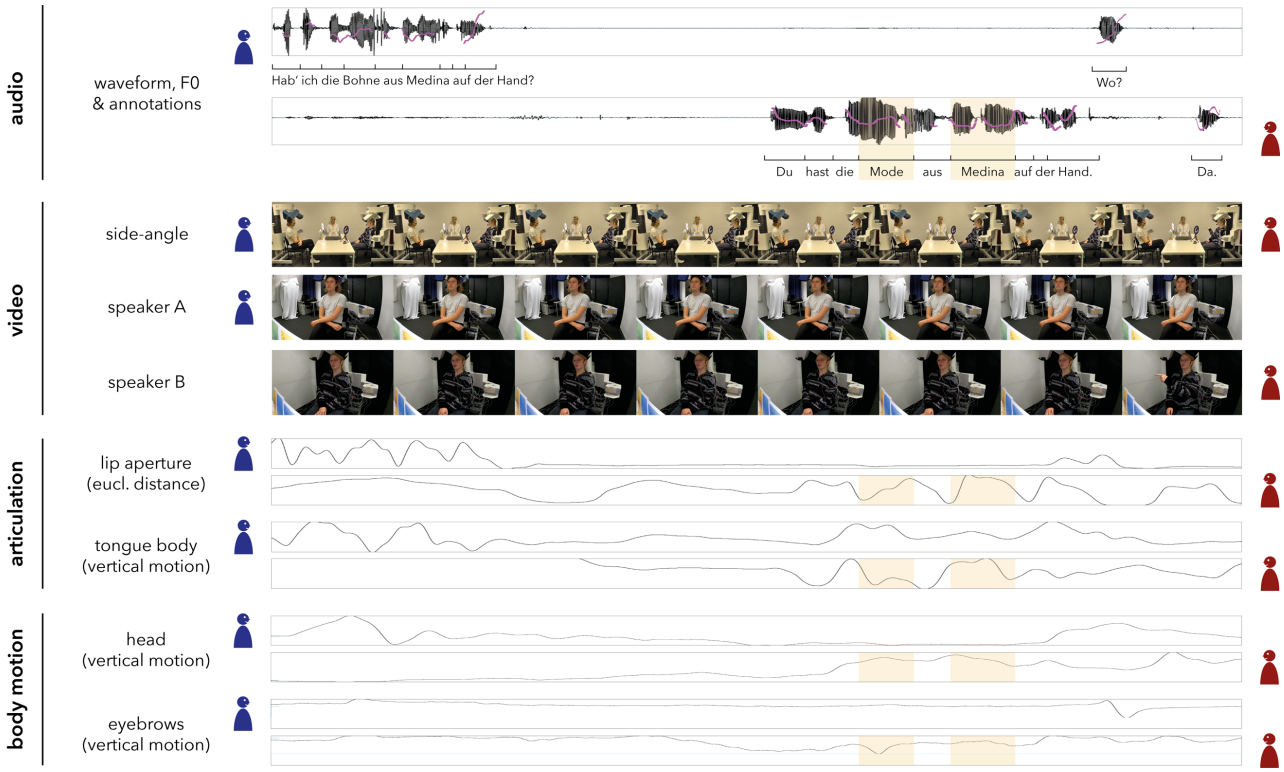


Figure 2: *Example of recorded multimodal data for selected parameters during one question-answer set by two speakers (blue and red). Target words are indicated by yellow rectangles. Based on the question (blue speaker), the answer (red speaker) has a focus structure with the object in corrective focus and the city in the background.*

Discussion. We present a novel methodological approach for multimodal recordings of two speakers in interaction, using dual 3D EMA, audio, and video simultaneously. We provide practical information on the procedure and technical setup, present benefits and potential pitfalls of using dual EMA in interactive multimodal recordings. Moreover, we introduce the cooperative card game *DiCE*, which elicits lexically and prosodically controlled speech material in an engaging task.

References

Babel, M., & Bulatov, D. (2012). The Role of Fundamental Frequency in Phonetic Accommodation. *Language and Speech*, *55*(2), 231–248. doi: 10.1016/j.jml.2011.07.004.

Cummins, F. (2002). On synchronous speech. *Acoustic Research Letters Online*, *3*(1), 7–11. doi: 10.1121/1.1416672.

Duran, N. D., & Fusaroli, R. (2017). Conversing with a devil's advocate: Interpersonal coordination in deception and disagreement. *PLoS ONE*, *12*(6), e0178140, 1–25. doi: 10.1371/journal.pone.0178140.

Lee, Y., Danner, S. G., Parrell, B., Lee, S., Goldstein, L., & Byrd, D. (2018). Articulatory, acoustic, and prosodic accommodation in a cooperative maze navigation task. *PLoS ONE*, *13*(8), e0201444, 1–26. doi: 10.1371/journal.pone.0201444.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*, 1404–1426. doi: 10.1111/j.1551-6709.2012.01269.x.

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, *66*, 249–264. doi: 10.1016/j.jml.2011.07.004.

Mukherjee, S., Legou, T., Lancia, L., Hilt, P., Tomassini, A., Fadiga, L., D'Ausilio, A., Badino, L., & Nguyen, N. (2018). Analyzing vocal tract movements during speech accommodation. *Proceedings of INTERSPEECH, 2-6 September, Hyderabad, India.*, 561–565. doi: 10.21437/Interspeech.2018-2084.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*, 132–142. doi: 10.1016/j.wocn.2010.12.007.

Oben, B., & Brône, G. (2016). Explaining interactive alignment: A multimodal and multifactorial account. *Journal of Pragmatics*, *104*, 32–51. doi: 10.1016/j.pragma.2016.07.002.

Tiede, M., Bundgaard-Nielsen, R., Kroos, C., Gibert, G., Attina, V., Kasisopa, B., Vatikiotis- Bateson, E., & Best, C. (2010). Speech articulator movements recorded from facing talkers using two electromagnetic articulometer systems simultaneously. *Proceedings of Meetings on Acoustics, 15-19 November, Cancun, Mexico*, art. 4pSC10. doi: 10.1121/1.4772388.

Tiede, M., & Mooshammer, C. (2013). Evidence for an articulatory component of phonetic convergence from dual electromagnetic articulometer observation of interacting talkers. *Proceedings of Meetings on Acoustics, 2-7 June, Montreal, Canada*, art. 3aSCa3. doi: 10.1121/1.4799497.

Auditory Targets for Sensory Feedback Control of Speech Change Over the Course of the Day

Frank H. Guenther[1], Alexander Acosta[1], Elaine Kearney[2]

[1]Boston University
[2]Queensland University of Technology
guenther@bu.edu, ajacosta@bu.edu, elaine.kearney@qut.edu.au

Introduction. It is widely believed that speech motor control involves idealized auditory targets corresponding to phonemes, syllables, and/or words of the native language. Perhaps the strongest evidence for this notion comes from experiments involving real-time perturbations of auditory feedback of a talker's own speech. For example, it has been repeatedly demonstrated that unpredictably perturbing pitch (or f0; e.g., Burnett et al., 1998) or formant frequencies (Tourville et al., 2008) leads to a *reflexive response* in the opposite direction of the perturbation, indicating that the speaker is attempting (usually subconsciously) to compensate for the perturbation so the production "sounds right". Since the auditory perturbation does not change somatic sensation, such experiments primarily engage the auditory feedback controller for speech, and the fact that speakers adjust their productions based on an induced auditory error indicates that they have an auditory expectation, or *auditory target*, for their speech movements.

It is usually assumed (often implicitly) that, at least in fluent adult speakers of a language, the auditory target for a sound remains essentially constant; that is, the speaker is attempting to achieve the same formants (say) for a vowel in the morning as in the afternoon or evening, and from one day to the next. It has been demonstrated that there is systematic variation in F1 and pitch for a given speaker saying a given vowel over the course of the day; on average, speakers tend to use higher pitch and F1 values later in the day compared to the morning (Heald & Nusbaum, 2015). This raises the question of whether these variations simply reflect changes in basic biophysical processes (such as changes in tissue volume or muscle stiffness/fatigue), or whether they reflect changes in the auditory targets utilized by the speech production mechanism. A change in pitch targets might not be surprising since changing baseline pitch does not affect phonemic identity, but a changing formant target would be surprising since vowel formants strongly affect vowel identity.

The current experiment investigated this issue by having subjects perform reflexive pitch and formant perturbation experiments in four different sessions at different times of day and on different days to capitalize on the natural variation in baseline pitch and F1 values reported by Heald & Nusbaum (2015). We then utilized simulations of the SimpleDIVA model (Kearney et al., 2022) fitting the resulting data to test between two hypotheses: (1) speakers utilize the same auditory target in different sessions although their productions vary in pitch/F1 across sessions (*fixed target model*), or (2) speakers utilize different targets at different times of the day and/or on different days (*variable target model*).

Methods. 23 native speakers of American English (ages 18-23; 18F, 5M) each participated in four experimental sessions at different times of the day on different days. Each session involved a total of 180 trials producing /CɛC/ words (e.g., "bed"), including 60 unperturbed trials, 30 F1 upshift trials, 30 F1 downshift trials, 30 f0 upshift trials, and 30 f0 downshift trials. F1 perturbations were applied throughout a shifted trial and involved a 30% change from baseline; f0 perturbations were 100 cents in magnitude and were applied within-utterance, with a randomly jittered onset (500-1000ms) relative to vocalization onset. For each participant and auditory parameter (f0, F1), the sessions with the highest (*high session*) and lowest (*low session*) baseline value of the relevant auditory parameter were selected for further analysis. For each participant, parameter, and condition (upshift, downshift, no shift), all trials for the condition were averaged together on a time-point by time-point basis; then the high session traces were averaged across subjects for each condition, as were the low session traces. For each session/condition, statistical tests were performed to determine whether participants significantly compensated for the perturbation. In order to test between the fixed and variable target models, two simulations of SimpleDIVA were performed to fit the averaged traces for a given session and condition: the first used the same target (estimated as the average pitch or F1 value of unperturbed trials across all sessions at each time point) and the second one used separate targets for the high session and low session (using the average pitch or F1 value from the unperturbed trials of that session only). Model fit quality was compared using the Akaike Information Criterion (AIC).

Results. Significant compensation ($p < 0.05$) was found for the group mean responses in both perturbed conditions of both sessions for both auditory parameters. Figure 1 illustrates the modeling results. Each panel shows, for a given condition (row) and model (column), the group mean trace (blue line), model fit (red line), and the auditory target used by the model (dashed line). Overall, the simulations strongly support the variable target model for both f0 and F1; comparison of AIC values indicated that the variable target model provided the superior fit in both cases ($p < .00001$). Qualitatively, this can be seen by the fact that the variable target model provides excellent fits for both shift directions in both sessions, whereas the single target model provides a poor fit for at least one direction/session.

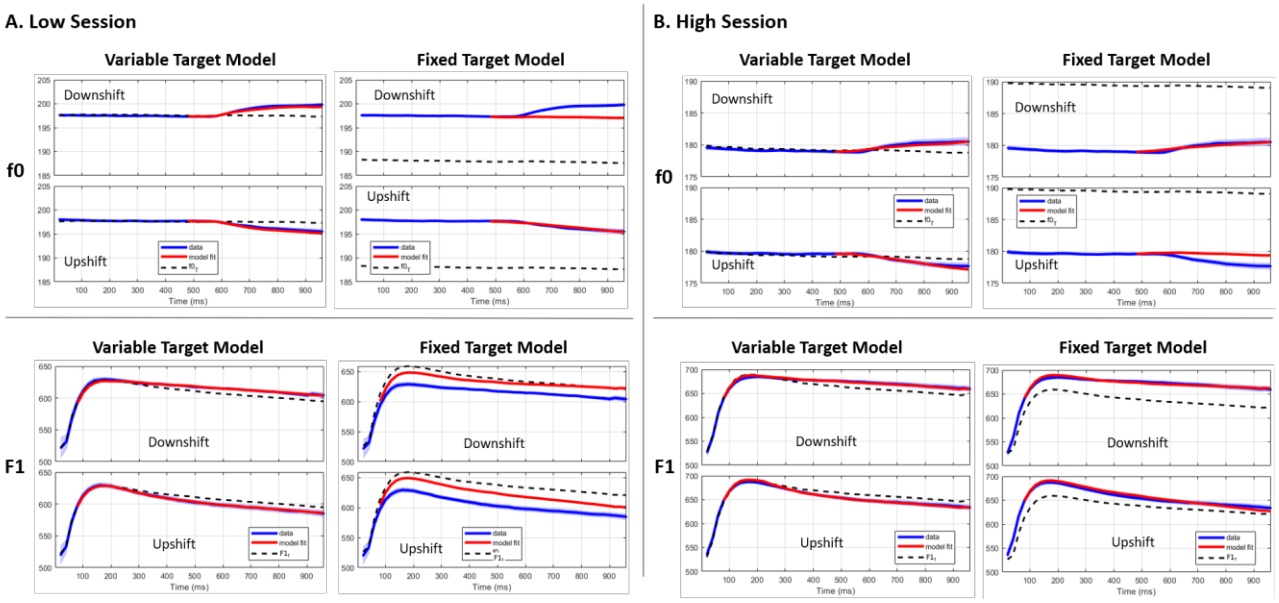


Figure 1: Experimental data (blue lines) and model fits (red lines) for (A) the low session and (B) the high session, with f0 traces in the top panels and F1 in the bottom panels.

Discussion. Our primary finding is that the target of the auditory feedback controller for speech (as revealed by unpredictable formant and pitch perturbations) is not a fixed "ideal" target, but instead varies over time, tracking natural variations in produced pitch and F1 that occur over time (Heald & Nusbaum, 2015). This finding has important implications for models of speech motor control, which typically assume (either implicitly or explicitly) that the auditory target for a speech sound remains essentially constant once well-learned (e.g., Guenther et al., 2006). The DIVA model (Guenther, 2016) currently posits that the auditory target for a speech sound (for example, a syllable with its own well-learned motor program) is encoded in projections from left ventral premotor cortex to higher-order auditory cortical areas; these "higher-level" targets act in parallel with "lower-level" targets from primary motor cortex to auditory cortical areas which encode the auditory targets corresponding to lower-level aspects of the utterance, such as phonemic gestures or muscle activation patterns. Within this framework, there are at least two possible interpretations of the current results. First, if projections from left vPMC to higher-order auditory areas are the primary source of the targets for auditory feedback control (as typically assumed in DIVA), then processing in left vPMC and/or these projections must change over the course of the day; that is, the motor program for a given phoneme, syllable, or word (presumed to reside in left vPMC) must be changing over the course of the day, even in adults. Another possibility is that the motor program itself does not change, but due to natural physiological variations over the course of the day, the same motor program in vPMC results in slightly different motor commands in bilateral ventral motor cortex (vMC), and the auditory target as probed by auditory perturbation experiments arises from vMC rather than from vPMC. Key questions for testing between these possibilities concern whether a speaker's perceptual "target" also changes over the course of the day, and if so, whether the changes in production targets track these perceptual changes. These questions are being tested in an ongoing follow-up study.

Summary. Our results unequivocally demonstrate that a talker's F1 target (as well as pitch target) for a vowel -- at least for the purposes of auditory feedback control -- is not constant even in a mature speaker, but instead changes systematically in concert with variations in produced formant frequencies over the course of the day. To our knowledge, no current model of speech production accounts for this variation; the implicit assumption of fixed auditory targets in adult speakers is widespread but in need of revision to account for these findings.

References

Burnett, T.A., Freedland, M.B., Larson, C.R., & Hain, T.C. (1998). Voice F0 responses to manipulations in pitch feedback. *Journal of the Acoustical Society of America*, 103(6), pp. 3153-3161.

Guenther, F. H. (2016). Neural control of speech. MIT Press.

Guenther, F.H., Ghosh, S.S., and Tourville, J.A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96, pp. 280-301.

Heald, S.L.M., & Nusbaum, H.C. (2015). Variability in vowel production within and between days. *PLOS ONE*, 10(9), e0136791.

Kearney, E., Nieto-Castañón, A., Falsini, R., Daliri, A., Heller Murray, E.S., Smith, D.J., & Guenther, F.H. (2022). Quantitatively characterizing reflexive responses to pitch perturbations. *Frontiers in Human Neuroscience*, 16, 929687.

Tourville, J.A., Reilly, K.J., and Guenther, F.H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage*, 39, pp. 1429-1443

From YIN to β-YIN: algorithm optimisation and performance analysis on auto-oscillating vocal folds replicas for normal and abnormal conditions

Raphaël Chottin[1], Xavier Pelorson[1], Didier Demolin[2], Annemie Van Hirtum[1]

[1] *CNRS-Université Grenoble Alpes/LEGI UMR 5519, Grenoble, France*

[2] *CNRS-Université Sorbonne-Nouvelle/LPP UMR 7648, Paris, France*

raphael.chottin@univ-grenoble-alpes.fr

Introduction. An algorithm for automatic detection and tracking of oscillations in physical signals is presented. It is based on YIN, a well known fundamental frequency estimator firstly introduced by De Cheveigné and Kawahara (2002) and known in speech sciences for its good performance in noisy conditions as shown by Sukhostat and Imamverdiyev (2015). An additional step including a parameter β is proposed and implemented, leading to a new algorithm called β-YIN firstly introduced by Chottin et al. (2023). A dataset of over 1500 physical signals of flow pressure, vocal fold displacement and acoustic pressure obtained from fluid-structure interaction experiments with mechanical vocal folds replicas, mimicking a normal or abnormal vocal fold structure, is used to evaluate the algorithm. Its two key parameters, σ and β, the first one being a parameter of the original YIN, are optimised for different signal to noise ratios (SNR).

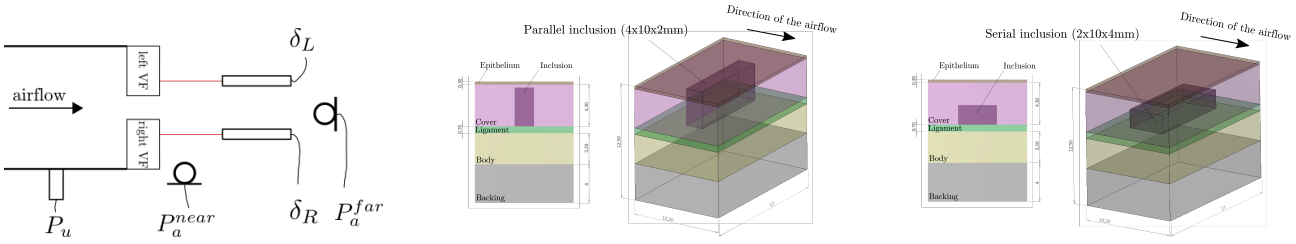
Methods. The β-YIN algorithm relies on YIN, a 6-step algorithm inspired by auto-correlation techniques (De Cheveigné and Kawahara 2002). While auto-correlation techniques aim at finding the maximum of an auto-correlation function to estimate the fundamental period of a signal s (step 1), YIN aims to minimise a difference function windowed over W (step 2) defined as $d_t(\tau) = \sum_{j=1}^{W}(s_j - s_{j+\tau})^2$. The function $d'_t(\tau)$ accounts for a normalisation of this difference function by its cumulative mean to prevent extreme low values to appear around the zero-lag. The resulting cumulative mean normalised difference function $d'_t(\tau)$ (step 3) is proportional to the aperiodic/total power ratio and defined as:

$$d'_t(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_t(\tau) \Big/ \left[\frac{1}{\tau}\sum_{j=1}^{\tau} d_t(j)\right] & \text{otherwise.} \end{cases} \quad (1)$$

A threshold σ is then introduced (step 4). The time lag associated to the first local minimum of $d'_t(\tau)$ with a value under σ is identified as the fundamental period T_0 resulting in frequency $f_0 = 1/T_0$. When no point is detected under the threshold σ, f_0 is set to 0, corresponding to the absence of periodicity. A parabolic interpolation is then performed around that local minimum to find the exact position of the local minimum (step 5). When performed over several overlapping windowed signal portions, a signal of f_0 as a function of time is obtained. Finally, the algorithm looks for the best local estimate at every time t (step 6). For each time t, the algorithm looks for a lower value of d'_θ for θ within a small interval in the vicinity of t. If a lower value is found, the frequency associated to θ is set to be the one associated to t.
Theoretically, if an oscillation starts during the acquisition of a physical signal, the first non-zero frequency point detected by YIN is the start of that oscillation, corresponding to the auto-oscillation onset which is a major quantity associated with voice production. In practice, it may happen that non-zero frequency points are detected when no oscillation is observable due to a complex oscillatory behaviour of the physical signal or an unfortunate effect of noise. In order to avoid to detect those events as the start of the oscillation, a 7th step is added to YIN, leading to β-YIN. In order to be detected as the time t_{onset} (resp. t_{offset}) of the start (resp. end) of the oscillation, the time t has to be followed (resp. preceded) by βW non-zero points with new parameter $\beta > 0$. Finding the onset and offset times t_{onset} and t_{offset} allows thus to find onset and offset frequencies f_{onset} and f_{offset} and to track any quantity characterising oscillation over time in a physical signal.

Results. β-YIN is evaluated over a large dataset of physical data obtained during fluid-structure interaction experiments. Flow-induced vibration of vocal folds (VF) is mimicked using deformable mechanical VFs replicas, *i.e.* composite silicone-molded replicas. Replicas are rectangular as in Van Hirtum et al. (2023) and composed of 5 layers of silicone with different mechanical properties (see Fig. 1). Some of the replicas have an inclusion, oriented either parallel or serial to the main auto-oscillation direction, embedded in the cover layer to mimic a pathological (or abnormal) vocal folds structure that tends to lead to complex vibration behaviours such a sub-harmonics generation or unstable frequency across

time (Van Hirtum et al. 2023). Two types of replicas (ML and MA) are used. The only difference is that the cover layer of the ML replicas has a higher Young's modulus.



(a) Experiment (b) VF replica with a parallel inclusion (c) VF replica with a serial inclusion

Figure 1: Schematics of the experiment (a) and silicone composite VF replicas with a parallel (b) and serial (c) inclusions.

β-YIN is evaluated over 321 samples using 5 signals: The upstream flow pressure P_u, the displacement signals $\delta_{R,(L)}$ of the right (and left) vocal fold and the acoustic pressures P_a^{near} in the near field and P_a^{far} in the far field. β-YIN has been evaluated over a window $W = 0.0154\ s$ corresponding to a minimum detectable frequency f_{min} = 65 Hz. First, the optimal value of σ minimising the error rate with respect to manually detected values on the signal's Hanning windowed spectrogram (resolution $\Delta f = 1.25$ Hz) is searched within the set $\{0, 0.1, 0.2, ..., 1\}$ for $\beta_{def} = 50$. β-YIN is considered to fail when either no detection is made or when the relative discrepancy, $\xi(y) = |y_{manual} - y_{YIN}|/y_{manual}$ with y being the onset (resp. offset) pressure and frequency P_{on}, f_{on} (resp. P_{off}, f_{off}), is greater than a defined tolerance of 10%. Once an optimal value σ_{opt} is found for β_{def}, the algorithm is applied for various values of β between 0 (no use of the βW condition) and 200. The optimisation of β can only be done after an optimisation of σ because a stable frequency estimate is needed for the last step in β-YIN. Obviously, manual detection is subjected to several biases. Overall optimised parameter sets (σ_{opt}, β_{opt}) for measured physical signals are shown in Table 1. The mean SNR is given as well.

Data type	P_{on}	f_{on}	P_{off}	f_{off}	mean SNR (dB)
P_u	9 % (0.5, 50)	14 % (0.5, 50)	12 % (0.5, 1)	12 % (0.5, 1)	18
δ_L	12 % (0.8, 5)	9 % (0.9, 5)	13 % (0.8, 5)	10 % (0.8, 5)	6
δ_R	12 % (0.9, 50)	12 % (0.8, 50)	12 % (0.9, 20)	8 % (0.8, 20)	7
P_a^{near}	21 % (0.8, 20)	24 % (0.9- 20)	25 % (0.9, 5)	19 % (0.9, 20)	10
P_a^{far}	35 % (0.9, 10)	50 % (0.9, 10)	57 % (0.9, 10)	49 % (0.9, 10)	4

Table 1: Percentage of β-YIN estimations with relative discrepancy ξ lower than 10% compared to the manually extracted value with, in brackets, the respective values of σ_{opt} and β_{opt} evaluated by β-YIN .

Noisier signals (SNR $\leq$ 7) such as δ_L or δ_R require a higher σ_{opt}. For less noisy signals (SNR $\geq$ 7), any value of σ between 0.2 and 0.9 leads to similar results. Values $10 \leq \beta \leq 50$ seem optimal, but didn't vary significantly for $\beta > 10$. Introducing β improved the performance with 20% up to 50%. The performance increases from about 10% to $\geq 20\%$ when considering acoustic pressures containing aperiodic as well as periodic noise, which decreases the performance.

Conclusion. β-YIN estimates onset and offset pressures and frequencies with a relative discrepancy ξ of less than 10% in up to 90% of the cases for the signals of SNR $\geq$ 18. It performs well for signals with lower SNR in the case of aperiodic noise but shows limitations when the noise shows periodic patterns. Next, a validation on signals measured on human speakers remains to be done as well as a more quantitative analysis of the perturbation observed in the physical signals.

References.

Chottin, Raphael, Ahmad Mohammad, Demolin Didier, and Pelorson Xavier (2023). "Objective detection and tracking of vocal folds auto-oscillation". In: *Proc. 10th Forum Acusticum.*

De Cheveigné, Alain and Hideki Kawahara (2002). "YIN, a fundamental frequency estimator for speech and music". In: *J. Acoust. Soc. Am.* 111.

Sukhostat, Lyudmila and Yadigar Imamverdiyev (2015). "A comparative analysis of pitch detection methods under the influence of different noise conditions". In: *J. of Voice* 29.

Van Hirtum, Annemie, Mohammad Ahmad, Raphaël Chottin, Oriol Guasch, and Xavier Pelorson (2023). "Experimental study of the influence of a rectangular vocal folds inclusion on their auto-oscillation". In: *Proc. 10th Forum Acusticum.*

Differential effect of phonemic contrast on corrective vowel production in child and adult speech

Melissa A. Redford[1], Carissa Diantoro[2]

[1]University of Oregon
[2]University of Oregon
redford@uoregon.edu, carissad@uoregon.edu

Introduction. When a listener mishears a speaker, the speaker may repeat exactly what they said before but this time under corrective (prosodic) focus. The specific adjustments associated with corrective focus have been described as localized hyperarticulation (de Jong, 1995). In the adult literature, hyperarticulation is very often supposed to represent a kind of goal maximization (Johnson et al., 1993; Lindblom, 1990) and has therefore been studied to infer something about the representations that guide speech articulation (e.g., Johnson et al., 1993; Schertz, 2013; Wedel et al., 2018). In clear speech studies, hyperarticulation is interpreted to suggest not only acoustic-perceptual speech motor goals but also that such goals reference discrete phonological representations of sound. For instance, Johnson and colleagues (1993) describe a perceptual 'hyperspace' effect, linking listener expectations for maximally distinct vowel productions to the speaker's speech motor goals in production. They then explicitly link such goals to discrete phonological representations, namely, phonemes. As phonemes are sound units of meaning contrast, the assumption of phonemes as speech motor goals entails that hyperarticulation result in contrast enhancement (Diehl, 2008; Lindblom, 1990). Yet, studies specifically designed to investigate the contrast enhancement hypothesis of hyperarticulation have focused on 2-way contrasts (Schertz, 2013; Wedel et al., 2018), making it possible to explain an increase in acoustic-perceptual distance between speech motor goals as due to exaggerated articulation rather than to true phonemically-motivated contrast enhancement. A more rigorous test of the contrast enhancement hypothesis is required to understand the relationship between speech motor goals and phonemes. The current study provides this test by adapting the design from a classic auditory feedback perturbation study (i.e., Houde & Jordan, 2002) to investigate the effect of the North American English 3-way lax vowel contrast on hyperarticulation in child and adult speech. Specifically, we used overt feedback from a listener to let the speaker know that their target /ɛ/-vowel word (e.g., *bet*) was being confused either with a minimal pair /ɪ/- or /æ/-vowel word (i.e., *bit* or *bat*). If children and adults target phonemes as speech motor goals, then they should correct the listener's misperception by adjusting /ɛ/ away from /ɪ/ and towards /æ/ when /ɛ/ is confused with /ɪ/ or away from /æ/ and towards /ɪ/ when it is confused with /æ/. If speakers do not make such adjustments and simply exaggerate /ɛ/ in the same way no matter the mishearing condition, then speech motor goals may be extralinguistic perceptual-motor targets rather linguistic ones—a possibility consistent with a theory of production that assumes holistic (exemplar-like) wordform representations (Redford, 2019; Davis & Redford, 2019).

Methods. Participants were 19 school-aged children (8-year-olds) and 30 young adults (18- to 22-year-olds). All were native American English speakers. All passed a pure tone hearing screening. Materials were CVC minimal triplet words with the English front lax vowels, /ɪ, ɛ, æ/. There were 10 triplets, resulting in 30 target words. Each word was paired with a picture for elicitation purposes. All picture–word correspondences were learned during a training phase that preceded the elicitation task. If a picture–word correspondence was forgotten during the task, the picture card was flipped over and shown to the participant as a written reminder of the correspondence.

The elicitation task took place in adjacent sound-attenuated experimental rooms, separated by a window. The experimenter and participant sat facing one another on either side of the window. This set up was used to enhance the plausibility of the mishearing manipulation. Elicitation occurred in three phases: a mapping phase, a mishearing phase, and a remapping phase. The experimenter used the pictures to elicit all 30 words twice in different fixed random orders during the mapping and remapping phase; only /ɛ/ words were elicited during the mishearing phase. The /ɛ/ words were 'misheard' by the experimenter as either the matched /ɪ/-word (/ɪ/ condition) or as the matched /æ/-word (/æ/ condition). During this phase of the experiment, the participant corrected the experimenter by repeating the /ɛ/-word in prosodic focus. The correction was done twice in a row on each trial (i.e., for each /ɛ/-word) so that each /ɛ/-word was elicited twice during the mishearing phase. As in auditory feedback perturbation studies, the specific mishearing manipulation was between subjects: half of the participants were assigned to the /ɪ/ condition and half to the /æ/ condition.

Participant speech was digitally recorded with a sampling rate of 44,100 Hz. A lavaliere microphone was attached to the participant's shirt to maintain a fixed mic-to-mouth distance. Acoustic segmentation and measurement were completed by a trained research assistant. Vowel identity was determined with reference to the intended CVC word, based on the fixed random order used during elicitation. Overall vowel duration was automatically extracted from the segmentations. F1, F2, and vowel intensity values were automatically extracted at 3 equal intervals around vowel midpoint. Normalized values of each measure were computed as follows: vowel duration, intensity, and formant values for each word within each phase within each participant were averaged; next, the averaged mapping phase values were

subtracted from the mishearing phase values (= in-focus) or from re-mapping phase values (= control) within word and participant.

Results. Linear mixed effects models tested for the fixed effects of corrective focus (in-focus vs. control) and mishearing condition (/ɪ/ versus /æ/) on the normalized acoustic measures while controlling for random effects of word and speaker. Analyses on children's speech showed clear effects of hyperarticulation when words were spoken under corrective focus: /ɛ/ was spoken with greater intensity [β = 6.32, SE = 0.75, p <.001], higher F1 [β = 76.91, SE = 11.46, p <.001], and higher F2 [β = 47.50, SE = 18.88, p <.001] when in-focus than when spoken normally during the remapping phase. There was no effect of mishearing condition on children's production nor any interaction between corrective focus and mishearing condition. By contrast, analyses of adults' speech indicated a significant interaction between corrective focus and mishearing condition on duration [β = 0.029, SE = 0.001, p <.001], intensity [β = 4.84, SE = 0.65, p <.001], and F1 [β = 22.68, SE = 5.63, p <.001]. Moreover, the direction of the interaction was consistent with contrast enhancement: adults produced longer, louder, and more open /ɛ/ when misheard as /ɪ/ but not as /æ/; that is, /ɛ/ moved away from /ɪ/ and towards /æ/ in the /ɪ/ mishearing condition. As in children's speech, corrected /ɛ/ was more fronted than normally-spoken /ɛ/ in adult's speech [β = 43.85, SE = 10.11, p <.001]. Figure 1 shows the results that distinguish child from adult speech.

Figure 1.

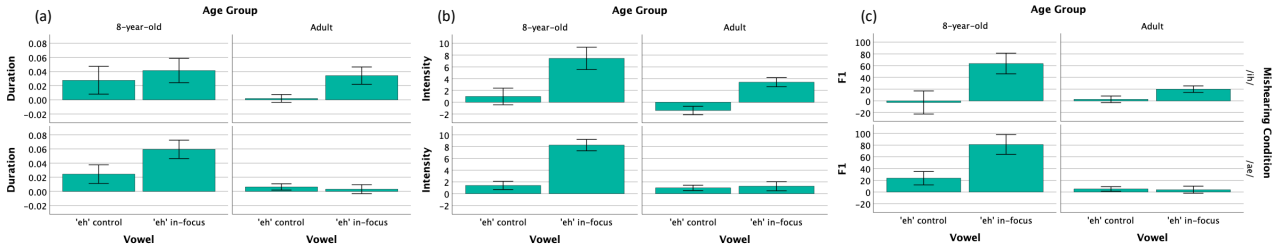


Figure 1: *Mean normalized duration (a), intensity (b), and F1 (c) is shown as a function of age group (child data in left-hand panels; adult data in right-hand panels) and mishearing condition (*/ɪ/ *condition in top panels;* /æ/ *condition in bottom panels). Error bars show the 95% confidence interval.*

Discussion. Overall, the results demonstrate an intriguing difference between child and adult speech that may bear on the maturation of speech motor goals or on age-related differences in the processing of other-produced speech or both. Whereas adults hyperarticulated /ɛ/ under corrective focus in a manner that might be described as contrast enhancing, children simply produced louder more open and more fronted /ɛ/ under corrective focus no matter the specific confusion they were correcting. It could be that children's speech motor goals are unrelated to phonemes. It could also be that, unlike adults, children were unable to conduct the phonemic analysis that would allow them to identify the specific difference between the sound they intended to produce (= /ɛ/) and the sound that listener's reportedly perceived (either /ɪ/ or /æ/). Although such an analysis does not require that the speaker target phonemes during speaking, it is at least consistent with theories that assume a direct relationship between phonemes and speech motor goals.

References

Davis, M., & Redford, M. A. (2019). The emergence of discrete perceptual-motor units in a production model that assumes holistic phonological representations. *Frontiers in Psychology*, 10, 2121.

De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, 97(1), 491-504.

Houde, J. F., & Jordan, M. I. (2002). Sensorimotor adaptation of speech I: compensation and adaptation. *Journal of Speech, Language, and Hearing Research*, 45(2), 295-311.

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, 505-528.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (eds.), *Speech production and speech modelling* (pp. 403-439). Dordrecht: Springer Netherlands.

Redford, M. A. (2019). Speech production from a developmental perspective. *Journal of Speech, Language, and Hearing Research*, 62(8S), 2946-2962.

Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics*, 41(3-4), 249-263.

Wedel, A., Nelson, N., & Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language*, 100, 61-88.

The contribution of voicing to coarticulatory nasalization in two varieties of English

Conceição Cunha[1], Jonathan Harrington[1], Philip Hoole[1]

[1]Institute for Phonetics, LMU Munich, Germany
cunha|jmh|hoole@phonetik.uni-muenchen.de

Introduction. American English VN sequences preceding voiceless consonants (e.g., 'sent') typically show more coarticulatory vowel nasalization and a reduced nasal consonant they when the receding consonant is voiced ('send', e.g., Beddor, 2009). Already Malécot (1960) reported an asymmetry in perception when the nasal consonant was extracted from the acoustic signal: American listeners could recover the nasal before voiced stop, but not when the stop was voiceless. In German, Carignan et al., (2021) showed a diminished the nasal gesture in size before voiceless in /Vntə/ (Ente, 'duck') than before voiced stops as in /Vndə/ (Ende, 'end'). Busà (2003, 2007) obtained comparable results for an Italian variety. Historically, the development of contrastive vowel nasalization is more likely to take place before voiceless than voiced consonants (Busà 2007; Hajek 1997), but it is not clear, what in speech physiology leads to these asymmetries.

Beddor has developed a model linking synchronic coarticulatory nasalization with the sound change known as vowel nasalization, arguing that the nasal gesture (velum) has a similar size across VNC contexts, but varies its alignment relative to the oral gesture (Beddor, 2009: 789). In this model, coarticulatory nasalization results from earlier alignment of the velum. In addition, after analyzing the onset of velum lowering relative to the vowel onset and the proportion vowel nasalization, Beddor (et al, 2018) showed that the onset of nasalization is earlier in the voiceless context when compared with voiced. Cunha and al. (submitted) tested Beddor's model with voiced /n, nd, nz/ contexts only by comparing American (USE) and Standard Southern British English (BRE). Overall, coarticulatory nasalization was greater in USE than in BRE. Vowels were indeed more nasalized, the nasal consonant less nasalized and the oral togue tip gesture for nasal /n/ was strongly lenited in USE when compared to BRE. The velum was stable in both varieties and not earlier in USE. Instead, the time of tongue tip raising peak velocity was close to the tongue tip maximum for USE, causing a shift in the acoustic boundary towards N (and lengthening the vowel), causing a greater overlap of the velum with the vowel and giving the illusion that the velum gesture aligns earlier in USE. For these voiced contexts, the suggestion is that the reduction of the tongue tip gesture causing coda reduction is the most important factor responsible for the increase of the vowel nasalization in USE.
In the light of the results from Cunha et al. (submitted), the main aim of the current paper is to extend their analyses to the voiceless context, by comparing /nt/ and /nd/ contexts in the same varieties of English. Two main predictions follow from this model, on the assumption that /nd/ shows less coarticulatory nasalization then /nt/. **i.** /nt/ has greater nasalization in the vowel and shorter nasal coda then /nd/, at least for USE; **ii.** The velum gesture has the same magnitude and temporal extent in both contexts, if vowel nasalization develops as a consequence of an earlier rephasing of a stable velum gesture.

Methods. Real-time magnetic resonance imaging (rt-MRI) data were acquired from 27 native speakers of standard Southern British English (SBE13 female) and 16 native speakers of US English (7 female). The US speakers were approximately equally distributed between Midland, Northeast, Southern, and West regions. The recording took place at the Max Planck Institute for Multidisciplinary Sciences in Göttingen, Germany. A 3-Tesla MRI system was used for image acquisition (Magnetom Prisma Fit, Siemens Healthineers, Erlangen, Germany) and an Optoacoustics FOMRI III fiber-optic dual-channel microphone (Optoacoustics Ltd) recorded audio simultaneously. The images were processed in Matlab. Every image in the dataset was first aligned to a reference image. After registration, a semi-polar grid consisting of 28 lines was applied semi-manually to the vocal tract, reaching from the glottis up to the alveolar ridge (see Carignan et al., 2021 for further details). For kinematic analysis of velum lowering, a ROI (region of interest) was manually defined around the spatial range encompassing the velum movements, which was then used as dimensions in principal component analysis (PCA). For the acoustics we have calculated the energy below 1kHz based on the YIN algorithm (de Cheveigne & Kawahara, 2002). The materials analyzed here consisted of 14 lexical words selected from a larger corpus (/nt/: bent, pent up, sent, bint, Pinter, sinter, bunt, punt, shunt; /nd/: band, feigned, fund, bend, binned). The words were spoken in the carrier phrase "saw <targetword> about two/four/five/six/ten", with narrow focus on the target word without repetitions.

Results. Fig. 1 (left panel) showing the displacement averaged between the peak velocities of velum movement. The main differences between /nt, nd/ trajectories were: (a) the magnitude of /nt/ is less (green usually less than gold) and (b) the lowering gesture for /nt/ is faster (steeper rise in the trajectory for green to the left of the peak) and (c) the /nt/ gesture is shorter. When testing the magnitude of velum, there was no effect of dialect and the displacement was found to be greater for /nd/ than for /nt/ only for the short vowels, not when the preceding vowel is /æ, eɪ/. These preliminary results so far (Fig. 1 left panel) show that the velum vesture is similarly aligned at acoustic vowel onset.

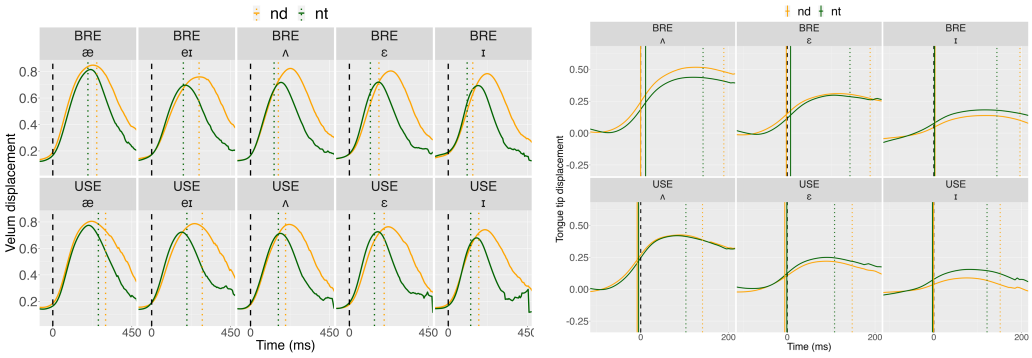


Figure 1: Left: *Displacement of the velum averaged aggregated by dialect, vowel, and /nd, nt/ context after alignment at the acoustic vowel onset (vertical dashed line at t = 0 ms). The green/gold vertical dashed lines for /nd, nt/ are at the times of the acoustic vowel offset. Right: Tongue tip displacement in three of these vowel contexts after aligned at the time of the peak tongue tip raising velocity.*

Fig. 1 (right panel) shows the tongue tip synchronised at peak tongue tip raising velocity (t = 0 ms, vertical dashed) with mean time of peak velum opening (solid) and mean time of peak tongue tip displacement (coloured, dotted). The tongue displacements are quite similar in both contexts and there is no evidence for a greater tongue tip undershoot in the /nt/ than in the /nd/ context. We then further analysed the data acoustically in order to determine whether the information for voicing in /n/ was diminished in /nt/ than in /nd/ which could mean that nasalization in the /nt/ coda is more difficult to identify than in /nd/. One of the cues for identifying nasalization acoustically is the presence of energy in the lower part (< 1 kHz) of the spectrum (e.g., Fujimura, 1962; House & Stevens, 1956). Fig 2 shows the dB-SPL level below 1kHz between the onset of /n/ (left vertical dashed line) and the time of the peak velocity of velum raising (dotted green/gold lines for voiced/voiceless). In all contexts and for both dialects, the energy is higher in the /n/ of /nd/ than in the /n/ of /nt/. This suggests that one of the driving forces for nasalization to diminish in the coda in /nC/ clusters where C is voiceless could be that there is a greater drop of the energy under 1 kHz for /nt/ in USE.

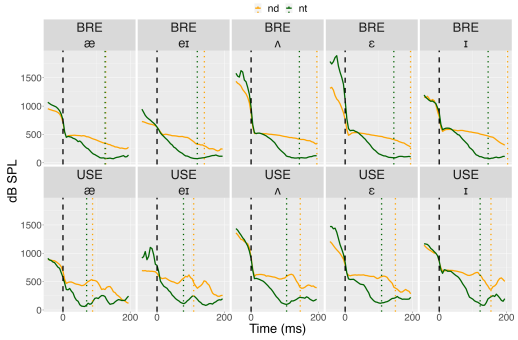


Figure 2: *Energy below 1kHz between the onset of /n/ (left vertical dashed line) and the time of the peak velocity of velum raising (dotted green/gold lines for voiced/voiceless)*

Discussion. So far there is little evidence for a leftwards shift of the velum gesture in /nt/ comparing with /nd/, but these analyses are still on progress. On the other hand, the acoustic analysis suggests that one of the driving forces for the diachronic waning of nasalization in a voiceless /nC/ context could be that /n/ is just more difficult to perceive when followed by a voiceless than voiced consonant (Ohala & Busà, 1995; Ohala & Ohala, 1991).

Selected references

Beddor, P., A. Brasher & Narayan, C. (2007). Applying perceptual methods to phonetic variation and sound change. In M.J. Solé et al. (Eds.), Experimental Approaches to Phonology. OUP: Oxford. (p.127-143).

Beddor, P., Coetzee, A., Styler, W., McGowan, K., and Boland E. (2018). The time course of individuals' perception of coarticulatory information is linked to their production: Implications for sound change. Language, 94, 931-968.

Carignan, C., Coretta, S., Frahm, J., Harrington, J., Hoole, P., Joseph, A., Kunay, E., and Voit, D. (2021). Planting the seed for sound change: Evidence from real-time MRI of velum kinematics in German. Language, 97.2, 333-364.

Cunha and al. (submitted). The physiological basis of the phonologization of vowel nasalization: a real-time MRI analysis of American and Southern British English: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4518960

Fujimura, O. (1962) Analysis of nasal consonants. Journal of the Acoustical Society of America, 34, 1865–75.

House, A., and Stevens, K. (1956). Analog studies of the nasalization of vowels. Journal of Speech and Hearing Disorders, 21, 218–32.

Malécot, A. (1960). Vowel nasality as a distinctive feature in American English, *Language* 36, 222–229

Ohala, M., and Ohala, J. (1991). Nasal epenthesis in Hindi. Phonetica, 48, 207– 220.

Ohala, J. & Busà, M. (1995). Nasal loss before voiceless fricatives: a perceptually¿based sound change. Rivista di Linguistica 7, 125-144.

The relation between breathing and utterance length in vocally learning birds

Susanne Fuchs[1], Lara S. Burchardt[12], Franz Goller[3]

[1]*Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin;* [2]Theoretische Neurophysiologie, HU zu Berlin;
[3]*Institut für Tierphysiologie, WWU Münster*
fuchs@leibniz-zas.de, lara.sophie.burchardt@hu-berlin.de, f.goller@utah.edu

Introduction. Birdsong and human speech share various neural and behavioral patterns that may provide us some hints about which capacities were requirements for language to evolve (Doupe and Kuhl 1999). Among them may be the capacity to prepare motor action for the upcoming speech, song or vocalization. For instance, it has been shown that humans roughly anticipate the length of the upcoming utterance in their respiration, with deeper and longer inhalations for longer utterances (Winkworth et al. 1995; Fuchs, Petrone, et al. 2013). On the other hand, the larynx can also compensate for the loss of air (Fuchs and Rochet-Capellan 2021), which has been found in longer sentence production. This compensation is reflected in a higher airflow resistance and voice quality changes (Zhang 2016; Aare et al. 2018). There have been recent attempts to investigate the preparatory actions in respiratory behavior and the laryngeal-respiratory interplay in other animals than humans. For example, Demartsev, Manser, and Tattersall (2022) tested thermal imaging cameras to obtain respiratory data in free-ranging meerkats. They found clear preparatory patterns in breathing but no correlation with utterance length. Riede, Schaefer, and Stein (2020) investigated a specific type of breath ("sigh", also called deep breath) in Sprague-Dawley rats (*Rattus norvegicus*) in vocal and non-vocal calls. They reported that changes in inhalation volume went hand in hand with changes in call duration. In a recent paper by Méndez, Dukes, and Cooper (2022) it has been shown that zebra finches (*Taeniopygia guttata*) and Bengalese finches (*Lonchura striata domestica*) prepare their respiratory action before song. The relation between utterance length and respiratory pressure was, however, negative, i.e., when birds prepared for longer utterances they started with a lower pressure and when they prepared for shorter utterances they started with a higher pressure. To what extent this is a matter of between-bird variation or also occurs within a bird was not discussed explicitly. The current paper aims to further explore the relation between breathing behavior and utterance length in vocal learning species for a better understanding of the evolutionary roots of this interplay. We define "utterance" as any unit of song which is uninterrupted by pauses and where air sac pressure does not drop below baseline.

Methods. Together 16 birds were recorded, four of the following species: zebra finch (*Taeniopygia guttata*), starling (*Sturnus vulgaris*), white-crowned sparrow (*Zonotrichia leucophrys*), canary (*Serinus canaria*). Ethical approval was provided by the Institutional Animal Care and Use Committee at the University of Utah. Air sac pressure was measured by surgically inserting a flexible cannula into an anterior thoracic air sac. A small hole was punctured through the body wall below the last rib, and the cannula was inserted and then secured with suture to the ribcage. The insertion site was also sealed with tissue adhesive to prevent leakage of air. The free end of the cannula was attached to a pressure transducer (Fujikura FPM-02PG), which was mounted on a backpack. After one day of recovery, air sac pressure and sound (Audio Technica AT 3032) were recorded for several days. Birds were kept in individual cages but were in acoustic contact with conspecifics to stimulate singing. In the zebra finch, directed songs were elicited by placing a female in a separate cage in front of the male. In the other species, playback of the bird's song was also occasionally used to induce singing. Data were recorded with Avisoft Recorder software in separate channels at 44.1 kHz. Pressure data were not calibrated and are therefore only used as relative changes within an individual.
From the available data, all zebra finches and two starlings have been annotated using Praat 6.4. (Boersma and Weenink 2023). The acoustic data of the zebra finches were relatively noisy. For this purpose, we obtained the on- and offset of their song based on high pass filtering of the pressure data (see Fig. 1 left, bottom signal). The filtered signal was then used to first automatically obtain silent and utterance intervals (window length=25ms, frequency range: 80Hz-10kHz, smoothing bandwidth=40Hz, and -20dB noise reduction), which were manually checked and corrected if needed. For the starlings, we used the acoustic signal (with better quality). Oscillations on the pressure signal would have been incomplete because these birds also produced voiceless sounds and clicks. The raw pressure signals were used to calculate inhalation

depth (i.e., the air sac pressure difference) and inhalation duration. The pressure data were first low-pass filtered and smoothed, eliminating the phonatory oscillations. In the next step, the first derivative (velocity) of the filtered pressure data was calculated (see Fig. 1 left, upper and middle signals) and its zero-crossings were used to identify the on- and offset of inhalation. Inhalation depth was calculated as the difference at the y-axis and inhalation duration as the difference between the two time points at the x-axis.

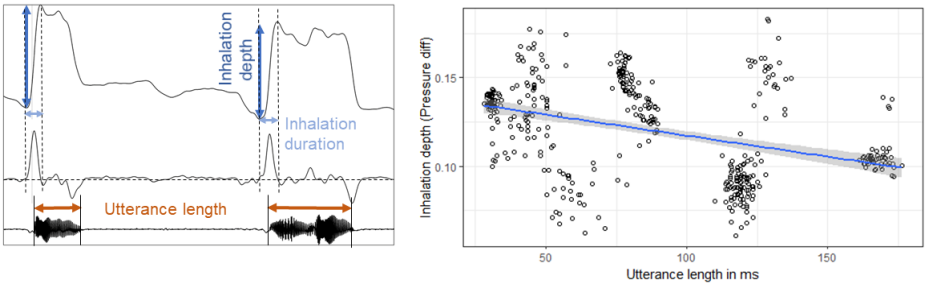


Figure 1: *Left: Annotation scheme with pressure data (top), first derivate (middle) and phonation (bottom). Right: Example of the negative relation between utterance length and inhalation depth for a zebra finch.*

Results and Discussion. For zebra finches (for each finch we have approximately between n=250-500 data points), results reveal a negative correlation between utterance length and inhalation depth (see Fig. 1 right). The longer the song, the less deeply the finches inhale. For the starlings, less data were available (for each starling n>44). Overall, no consistent relation was found. Each starling had some files where inhalation depth either did not vary to the length of the song or showed a positive or negative relationship. Further analysis is needed to differentiate between different songs, individuals, and species. The different patterns might be specific to selected songs and individuals. Concerning inhalation duration, we found for most birds a positive relationship to peak inhalation (offset), i.e., the longer the inhalation duration, the higher the inhalation peak. Our results so far are in agreement with earlier work by Méndez, Dukes, and Cooper (2022) on zebra finches, which showed preparatory actions in breathing behavior. The negative relation is, however, different from human animals and deserves further explanations. We noticed for shorter songs that zebra finches started with a higher pressure, but this pressure dropped rapidly. In longer utterances the pressure started at a lower level but was kept relatively stable over time, speaking for a fine-tuned interaction of the respiratory system with the syrinx in birds. This may show some specificities of bird song Goller (2022) in comparison to human speech. Our preliminary findings reveal that human animals and zebra finches use anticipatory mechanisms to prepare for the upcoming vocalization, but they use this in different ways.

References.

Aare, Kätlin, Pärtel Lippus, Marcin Wlodarczak, and Mattias Heldner (2018). "Creak in the respiratory cycle". In: *Proc. Interspeech 2018*, pp. 1408–1412. DOI: 10.21437/Interspeech.2018-2165.

Boersma, Paul and David Weenink (2023). "Praat: doing phonetics by computer [Computer program]". In: *http://www. praat. org/*.

Demartsev, Vlad, Marta B Manser, and Glenn J Tattersall (2022). "Vocalization-associated respiration patterns: thermography-based monitoring and detection of preparation for calling". In: *Journal of Experimental Biology* 225.5, jeb243474.

Doupe, Allison J and Patricia K Kuhl (1999). "Birdsong and human speech: common themes and mechanisms". In: *Annual review of neuroscience* 22.1, pp. 567–631.

Fuchs, Susanne, Caterina Petrone, Jelena Krivokapić, and Philip Hoole (2013). "Acoustic and respiratory evidence for utterance planning in German". In: *Journal of Phonetics* 41.1, pp. 29–47.

Fuchs, Susanne and Amélie Rochet-Capellan (2021). "The respiratory foundations of spoken language". In: *Annual Review of Linguistics* 7, pp. 13–30.

Goller, Franz (2022). "The syrinx". In: *Current Biology* 32.20, R1095–R1100.

Méndez, Jorge M, Jacqueline Dukes, and Brenton G Cooper (2022). "Preparing to sing: respiratory patterns underlying motor readiness for song". In: *Journal of Neurophysiology* 128.6, pp. 1646–1662.

Riede, Tobias, Charles Schaefer, and Amy Stein (2020). "Role of deep breaths in ultrasonic vocal production of Sprague-Dawley rats". In: *Journal of Neurophysiology* 123.3, pp. 966–979.

Winkworth, Alison L, Pamela J Davis, Roger D Adams, and Elizabeth Ellis (1995). "Breathing patterns during spontaneous speech". In: *Journal of Speech, Language, and Hearing Research* 38.1, pp. 124–144.

Zhang, Zhaoyan (2016). "Respiratory laryngeal coordination in airflow conservation and reduction of respiratory effort of phonation". In: *Journal of Voice* 30.6, 760–e7.

Attentional demand on speech processing: evidence from dual-task interference on vowel space and V-to-V anticipatory coarticulation according to task properties

Michaela Pernon[1,2], Daria D'Alessandro[1,3]

[1] Laboratoire de Phonétique et Phonologie, CNRS & Université Sorbonne Nouvelle, Paris, France.
[2] Neurology Department, Hôpital Fondation A. de Rothschild, Paris, France.
[3] University of Washington, Seattle, WA, USA.
michaela.pernon@gmail.com, dariada@uw.edu

Introduction. Speech is usually fast, accurate, and automatic, but in dual-task situations, which are common in everyday life - for instance talking while driving - it may be less straightforward. Observations from dual-task paradigms in experimental settings show that performing two tasks simultaneously can lead to interference from one task on the other, compared to when the tasks are done in isolation. Under dual-task conditions, the change in speech parameters is interpreted as a marker of an additional attentional demand on the speech production system.
Little is known about the effect of dual-task on speech processing levels. Indeed, the results on dual-task effects on speech production are controversial or difficult to generalize for two main reasons. First, the properties of the speech tasks and the concurrent tasks vary across studies in terms of the degree of attentional demand, automaticity, modalities or mode of presentation of the stimuli of non-verbal tasks, etc. Second, various speech parameters have been studied in neurotypical speakers, spanning from f0 (Fuchs et al., 2015) to voice intensity (Dromey & Bates, 2005), to temporal measures at the utterance level such as speech initiation time, pause time (Ho et al., 2002), speech rate (Kemper et al., 2010). The few acoustic studies that have tackled the issues on the interference of dual-task at the segmental level found no dual-task effect on vowel space area (Whitfield et al., 2019), and a reduction in F1 and F2 slopes in diphthong transitions (Dromey et al., 2010). To our knowledge, there is no data on the effect of dual-task on anticipatory V-to-V coarticulation. Since anticipatory coarticulation can be considered as a cue of planning of the movements for the production of forthcoming speech units (e.g., Whalen, 1990), the question of whether it can be affected by increased attentional demand, common in everyday life situations, is of particular interest. It could also help clarify the modelling of an additional attentional demand and its allocation at the level of speech processing. Indeed, it raises the question of the place of the attention in the speech production system, i.e. at which processing levels, such as planning, programming, or execution, it intervenes.
The main aim of this study is to investigate the effect of different dual-task conditions on anticipatory V-to-V coarticulation and vowel space, in order to shed a light on the impact of additional attentional demand on speech processing levels from the segmental cues of speech production. The second aim is to identify which properties of the non-verbal tasks (varying in degree of attentional demand and mode of presentation of the stimuli) might influence the dual-task effect.

Methods. 27 young adults (6 M, 21 F), aged 19 to 29 years old (M: 22, SD: 2,8), were asked to repeat the French sentence: "*papa et papi papotaient tout à coup*", /papaepapipapɔtɛtutaku/ (*Dad and grandpa were suddenly chatting)* for 55'' under different conditions in a dual-task paradigm. Non-verbal tasks properties were varied for: (1) the degree of attentional demand - a go task (GO) involving sustained and selective attention vs. a go-nogo task (GONOGO) also involving inhibition -, and (2) the mode of presentation of the stimuli - continuous versus discrete -. In the continuous mode, stimuli were presented simultaneously on a paper sheet (paper-and-pencil visuo-motor task), while in the discrete mode, stimuli appeared one by one on a computer screen (computerized visuo-manual task). The speech task was produced in a speech-only condition at the beginning and at the end (SINGLE) to control for a potential learning bias (and averaged), and under dual-task simultaneously with 4 non-verbal tasks (DUAL). The non-verbal tasks were performed first in a single condition. The order of the modes, and of the non-verbal tasks was counterbalanced across the 27 participants.
For the spectral analyses of the speech task, F1 and F2 of V1 /a/ of /papa/, V1 /a/ and V2 /i/ of /papi/, and /u/ of /ku/ were extracted. Two vowel space metrics were computed in Bark for each of the cardinal vowels /a/ (/papa/), /i/, /u/: (1) the vowel distance to centroid (DistCentroid), computing here the average distance of the 3 vowels' centroids to the overall centroid of the speaker's vowel space, indication of greater or lesser centralisation, and (2) the distance to category centroid (V-Dispersion) of each of these token as a measure of dispersion of its category (Audibert et al., 2015). V-to-V anticipatory coarticulation was analysed on the word /papi/ and considered as the degree of acoustic assimilation between V1/a/ and V2/i/, using the measure: ((F2 - F1 /a/) – (F2 - F1 /i/)) / (F2 – F1 /i/) (Coa-Index).
To estimate the influence of mode and attentional demand, a dual-task effect index (DTE) was computed for all speech metrics as: (SINGLE – DUAL (GO or GONOGO)) / SINGLE. Finally, the duration of each vowel was taken to test the

influence of temporal parameters on the spectral ones. Linear mixed-effects models were used for the analyses of dual-task effect.

Results. Our results show an effect of the dual task on V-to-V coarticulation in both the discrete ($F(2, 1909) = 5.17$, $p = .005$) and in the continuous mode: ($F(2, 1971) = 12.01$, $p < .001$) and on the vowel distance to centroid in the continuous mode ($F(2, 13) = 8.19$, $p = .003$). There was no effect of dual task on vowel dispersion.
In particular, it is worth noticing that coarticulation increased in all DUAL conditions. However, while the degree of attentional demand of the non-verbal tasks did not influence the dual-task effect, the mode of presentation of the stimuli of the non-verbal tasks affected coarticulation degree, with a stronger effect of the dual-task in continuous mode. The change in coarticulation degree was not affected by changes in vowel duration.
Vowel space area was less affected by dual-task interference with only a centralisation of high vowels in continuous mode, with an increase in centralisation at the decreasing of vowel duration ($F(1,13) = 7.66$, $p = .005$).

Discussion. In contrast to the results of previous studies on segmental parameters, which found no dual-task effects in neurotypical speakers (Dromey et al., 2010; Whitfield et al., 2019), our findings revealed robust dual-task effects for anticipatory V-to-V coarticulation and, to a lesser extent, for vowel space area. The increased coarticulation in the dual-task condition could be interpreted as resulting from motor economy (Patri et al., 2015) or hypospeech (Lindblom, 1990). While a hypoarticulation in the DUAL condition could be supported by the partial vowel centralization here found, this hypothesis needs to be confirmed by analysing the performance of the non-verbal tasks in a bidirectional analysis, in order to infer the strategies used by the speakers.
Through the changes in segmental cues under dual-task conditions, these results also question the impact of an attentional demand on speech processing levels, as briefly addressed by some models. These models relate it to some kind of monitoring in sensory feedbacks (FL-F model: Van der Merwe, 2021) or in forward prediction mechanisms and sensorimotor inhibitory activity (HSFC: Hickok, 2014). Additional attentional demand would affect here all speech processing levels, probably more planning than programming / execution, given the large coarticulation effects.
In conclusion, our results highlight the significance of considering the mode of presentation of stimuli of non-verbal tasks when examining dual-task effects on speech processing. Finally, in this study we only analysed the speech of 29 young adult participants. Further research including a larger sample size and a population more varied in age could allow generalisation of these findings to older age groups or to speakers with speech disorders, with a view to clinical application. Furthermore, correlations with dual-task changes in an extended set of segmental parameters, also in other speech dimensions, would be of interest in order to have a global view of the impact of attentional demand on the speech production system.

References

Audibert, N., Fougeron, C., Gendrot, C., & Adda-Decker, M. (2015). Duration-vs. style-dependent variation: a multiparametric investigation. In *Proceedings of 18th International Congress of Phonetic Sciences (ICPhS)*.

Dromey, C., & Bates, E. (2005). Speech interactions with linguistic, cognitive, and visuomotor tasks. *Journal of Speech, Language, and Hearing Research*, *48*(2), 295-305.

Dromey, C., Jarvis, E., Sondrup, S., Nissen, S., Foreman, K. B., & Dibble, L. E. (2010). Bidirectional interference between speech and postural stability in individuals with Parkinson's disease. *International journal of speech-language pathology*, *12*(5), 446-454.

Fuchs, S., Reichel, U. D., & Rochet-Capellan, A. (2015). Changes in speech and breathing rate while speaking and biking. In *Proceedings of 18th International Congress of Phonetic Sciences (ICPhS)*.

Hickok, G. (2014). Towards an integrated psycholinguistic, neurolinguistic, sensorimotor framework for speech production. *Language, Cognition and Neuroscience*, *29*(1), 52-59.

Ho, A. K., Iansek, R., & Bradshaw, J. L. (2002). The effect of a concurrent task on Parkinsonian speech. Journal of Clinica1 and Experimental Neuropsychology, 24(1), 36-47.

Kemper, S., Schmalzried, R., Hoffman, L., & Herman, R. (2010). Aging and the vulnerability of speech to dual task demands. *Psychology and aging*, *25*(4), 949.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Speech production and speech modelling (pp. 403-439). Dordrecht: Springer Netherlands.

Patri, J. F., Diard, J., & Perrier, P. (2015). Optimal speech motor control and token-to-token variability: a Bayesian modeling approach. *Biological cybernetics*, *109*, 611-626.

Van Der Merwe, A. (2021). New perspectives on speech motor planning and programming in the context of the four-level model and its implications for understanding the pathophysiology underlying apraxia of speech and other motor speech disorders. *Aphasiology*, *35*(4), 397-423.

Whalen, D. H. (1990). Coarticulation is largely planned. Journal of Phonetics, 18(1), 3-35.

Whitfield, J. A., Kriegel, Z., Fullenkamp, A. M., & Mehta, D. D. (2019). Effects of concurrent manual task performance on connected speech acoustics in individuals with Parkinson disease. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2099-2117.

Examining Speech Perception of Non-Errored Pronunciations in Children with Speech Sound Disorders

Elaine R. Hitchcock[1], Laura L. Koenig[2,3]

[1]Montclair State University
[2]Adelphi University
[3]Haskins Laboratories

hitchcocke@montclair.edu, lkoenig@adelphi.edu, laura.koenig@yale.edu

Introduction. Previous studies assessing speech perception in children with speech sound disorder (SSD) suggest a) inconsistent, if any, differences from typically-developing peers (TD) and/or b) that children with SSD perceive inaccurate productions as acceptable variants of their distorted or misarticulated speech productions (Lof & Synan, 1997; Shuster, 1998). Thus, comparisons of TD and SSD perception may depend on whether or not the sounds being assessed are produced accurately or in error by the child (Locke, 1980) as well as variations in the tasks or stimuli (e.g., synthetic speech, synthetically-altered natural speech, and natural speech). Much work assessing children's speech perception has used synthetic speech, following classic studies such as Kuhl & Miller (1978); however, extending findings to natural speech is not straightforward. Perceptual judgments may also be influenced by distributional properties of the dataset (Hitchcock & Koenig, 2021; Maxwell & Weismer, 1982). The primary aim of the present work is to investigate whether TD children and those with SSD differ in their perceptual labeling of stop-initial words in young children.

Methods and Analysis. Listening participants included 15 monolingual English-speaking typically-developing (TD) children (9F, 6 M; age range 6;0–10;6 and 14 monolingual English-speaking children diagnosed with a speech sound disorder (SSD; 6F, 8M; age range 6;10–10;5). All children demonstrated typical language functioning status, hearing sensitivity within normal limits, age-appropriate cognitive and motor milestones, and no significant medical or psychological history. Gender and ethnicity were not controlled. None of the children with SSD were perceived to have any voicing errors. Listeners were asked to perform a forced-choice identification task in response to child stimuli blocked by place of articulation (POA). All participants completed one data collection session of approximately 60–90 minutes conducted in a WhisperRoom MDL 10284 S sound booth. Stimuli were presented via Dell Latitude E6500 computers using a SB1700 soundcard and Sennheiser HD280 headphones. Stimuli consisted of a subset of single word targets from Hitchcock and Koenig (2013). Two-year old children spontaneously produced the CV target words "*boo", "pooh",* or "*doe", "toe"* in response to pictured stimuli. Voice onset time (VOT; Lisker & Abramson, 1964) was measured using a Pentax Computerized Speech Lab (Model-4500), referencing the acoustic waveform and wideband spectrogram. From this dataset of four words, six exemplars were chosen from each of six children with short-lag /b d/, short-lag /p t/, long-lag /b d/, and long-lag /p t/ targets. For each POA and VOT category, /b d/ and /p t/ VOTs were bimodally distributed (shorter for voiced targets), separated by a 5 ms gap. The bimodal VOT distribution consisted of four VOT ranges: Appropriate for /b d/ (0–10 ms), appropriate for /p t/ (67.5–100 ms), inappropriate for /b d/ (25–62.5 ms), and inappropriate for /p t/ (15–25 ms) (see Figure 1). Word and vowel context was restricted to the CV target words. Listener ratings were considered accurate if they matched the child's target.

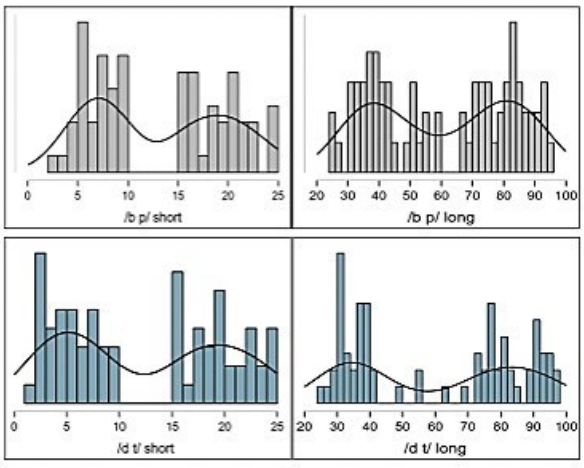


Figure 1: *Distribution of stimuli along the VOT continuum*

Results. Listener responses are organized using the four categories defined above: (1) *Appropriate VOTs*: Productions of /p t/ with long-lag VOTs and productions of /b d/ with short-lag VOTs. (2) *Inappropriate VOTs*: Productions of /p t/ with short-lag VOTs and productions of /b d/ with long-lag VOTs are presented in **Table 1**.

Significant results from Shapiro-Wilks tests indicated deviation from normality for all comparisons; thus, Mann Whitney U tests were calculated to assess group differences for all VOT categories. Group differences were only significant for one comparison (long-lag /b/). This could suggest largely comparable speech perception for TD and SSD children. Interestingly, however, for three of the four inappropriate VOT categories, accuracy was actually higher for those with SSD (albeit not always rising to the level of significance).

In all cases, accuracy was much higher for appropriate VOTs than for inappropriate VOTs, suggesting that listener judgments were mainly driven mainly by VOT. Higher accuracy in both groups for long-lag /d/ could reflect secondary cues available in the stimuli.

Table 1: *Accuracy (Acc) and standard deviation (SD) by group for appropriate and inappropriate VOT categories. TD=Typically developing, SSD=Speech sound disorder. Mann–Whitney U test results pairwise comparison by population. Values were judged to be statistically significant using a standard criterion p = .05*

	Appropriate VOTs				Inappropriate VOTs			
Group	Short-lag		Long-lag		Short-lag		Long-lag	
	b	d	p	t	p	t	b	d
TD Acc	90%	88%	97%	97%	30%	38%	33%	64%
TD SD	30.27	33.35	16.44	15.91	45.70	48.48	47.05	48.13
SSD Acc	88%	88%	95%	97%	34%	43%	39%	63%
SSD SD	32.88	32.88	20.89	17.55	47.46	49.54	48.80	48.19
Mann-Whitney U	133200	135468	133650	135288	129960	128916	128016	135792
p-value	0.279	0.824	0.124	0.574	0.119	0.083	*0.046	0.943

**Indicates statistical significance (p <.05).*

Discussion. In both child groups, labeling was highly accurate for targets with *appropriate* VOTs. This is consistent with work showing high accuracy in adults for child targets with appropriate VOT values (Hitchcock & Koenig, 2021). The statistical results are also consistent with studies suggesting that children with SSD do not show clear perceptual deficits on non-errored sounds compared to their TD peers. At the same time, slightly higher accuracy levels for *inappropriate* VOT targets in children with SSD warrants further investigation. Potentially, children with SSD could have coarser perceptual skills and wider boundaries in their categorical labeling functions than TD children, even for sounds that are not produced in error.

Hitchcock and Koenig (2021) explored adult labeling of toddler speech that did not incorporate the bimodal stimulus distributions used here. Adult responses there to inappropriate VOT values for /p t/ were considerably lower than observed here (11–15%). In follow-up studies, we have observed higher labeling accuracy for adults and children listening to bimodally-distributed data. This suggests that bimodal distributions of VOT within short- and long-lag ranges may lead to higher-than-expected accuracy for listener responses. Future research should also explore how secondary cues may contribute to perceptual judgments, comparing children and adults, and TD vs. SSD child groups.

References

Hitchcock, E. R., & Koenig, L. L. (2013). The effects of data reduction in determining the schedule of voicing acquisition in young children. *Journal of Speech, Language, and Hearing Research,* 56(2), 441–457.

Hitchcock, E. R., & Koenig, L. L. (2021). Adult perception of stop consonant voicing in American-English-learning toddlers: Voice onset time and secondary cues. *Journal of the Acoustical Society of America*, *150*(1), 460–477.

Kuhl, P.K., & Miller, J.D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America,* 63(3), 905–917.

Lisker, L., & Abramson, A. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word,* 20(3), 384–422.

Locke, J. (1908). The inference of speech perception in the phonologically disordered child. Part I: A rationale, some criteria, the conventional tests. *Journal of Speech and Hearing Disorders, 45*(4), 431–444.

Lof, G. L., & Synan, S. T. (1997). Is there a speech discrimination/perception link to disordered articulation and phonology? A review of 80 years of literature. *Contemporary Issues in Communication Science and Disorders, 24(*Spring), 57–71.

Maxwell, E. M., and Weismer, G. (1982). "The contribution of phonological, acoustic, and perceptual techniques to the characterization of a misarticulating child's voice contrast for stops," *Applied Psycholinguistics,* 3(1), 29–43.

Shuster, L. I. (1998). The perception of correctly and incorrectly produced /r/. *Journal of Speech, Language, and Hearing Research ,* 41(4), 941–950.

Auditory feedback of speech: comparison between aerial and bone-conducted pathways

Raphaël Vancheri[1], Coriandre Vilain[1], Nathalie Henrich Bernardoni[1], Pierre Baraduc[1]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble-INP — GIPSA-lab*
38000 Grenoble, France
pierre.baraduc@gipsa-lab.fr

Introduction. During speech, sound waves produced in the vocal tract propagate not only through the air, but also through soft tissues and bones. They are transmitted to the cochlea, either directly through the parietal bone, or indirectly through the motion of the eardrum and ossicles. Because of the experimental difficulty to access this internally-conducted signal, little is known about this part of the auditory feedback in the perception and control of one's own speech. Internal conduction (henceforth named "bone conduction", as customary in this field) was studied in animals (e.g. Tonndorf 1966) or in cadavers (Eeg-Olofsson et al. 2008), and the biophysical processes of this transmission were characterized (as reviewed in Stenfelt and Goode 2005). Since Békésy (1949), who first reported that the aerial-conducted (AC) and bone-conducted (BC) speech signals were of similar magnitude, few studies have investigated the BC feedback of speech. Pörschmann (2000) used a masking procedure on isolated phonemes to study the spectral features of the BC speech, and reported a difference between voiced and unvoiced sounds. Later, Reinfeldt et al. (2010) used the direct recording of ear canal vibrations as a proxy for the bone-conducted signal in order to describe the differences in spectral signature between AC and BC speech for several isolated consonant and vowels. To delve deeper and study BC in natural speech utterances, we adopted the same method, notwithstanding its limitations (it does not estimate BC sounds above 4 kHz).
Here we chose to focus on a comparison between the *informational content* of the aerial and bone-conducted components of the auditory feedback, besides their perceptual differences. To this end, we used a voice conversion procedure to evidence moments when one component cannot successfully predict (that is, be converted into) the other. This allows to find phonemes for which AC and BC components bear a different information.

Methods. Six French-speaking subjects participated (4 women, 2 men, 22-49 yr). They pressed their left ear against an "earbox", an in-house developed large earmuff allowing to isolate acoustically (by > 30 dB) one ear while preventing any occlusion effect (Figure 1A). A silicon probe microphone recorded the sound inside the ear canal, close to the eardrum, to estimate the BC component. A second microphone recorded the AC sound next to the contralateral (free) ear. Subjects were required to read aloud the first 100 sentences of the FHarvard corpus.
Both recordings were denoised through spectral subtraction. Phonetic segmentation was then carried out with Praat using the EasyAlign module; results were checked and corrected if necessary. Voice conversion was achieved with the method of Toda and Shikano (2005) as implemented by T. Hueber, who kindly shared his code (Hueber and Bailly 2016). To summarize, signals were transformed into time series of mel-cepstral coefficients with the Speech Processing Toolkit (SPTK). For each conversion (AC $\rightarrow$ BC or conversely), a Gaussian mixture model was trained. Finally, a Gaussian mixture regression allowed the conversion across mel-cepstral coefficients.Training was done on 80 sentences, and prediction on the remaining 20 ones.
For each signal sample, the norm of the difference between the spectral envelope of signal X and of its prediction from signal Y was computed on the $0-4$ kHz band, and considered as a measure of the informational content specific to X (inconvertible from Y), with X and Y being AC and BC or conversely.

Results. Figure 1B summarizes our observations on the phonemes of the recorded corpus. The AC component of open vowels like /a/, /ɔ or /ɛ/ carries information absent from the BC component. Reciprocally, in the nasal vowels, the BC component carries information absent from AC, as could be expected; this also happens during the release of occlusive consonants, which we did not anticipate. Most interestingly, closed vowels (/i/), nasal consonants (/m/, /n/) and fricatives like /ʃ/, /ʒ/ or /s/ seem to have a specific but different information in both AC and BC signals.

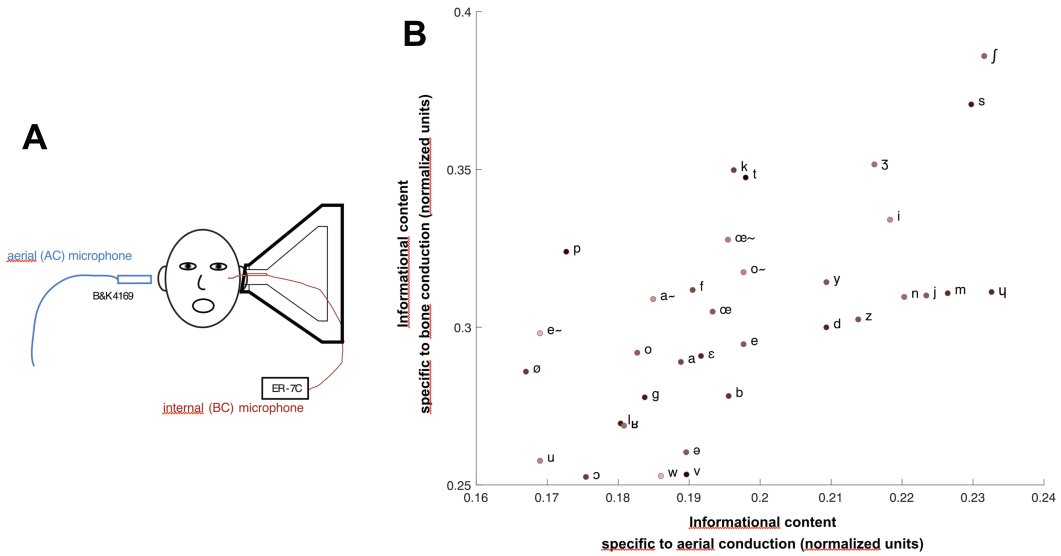


Figure 1: A. Experimental setup. B. Difference in information carried by AC and BC components, as estimated from the voice conversion method, for each phoneme of the recorded corpus.

Discussion. To our knowledge, these results are the first attempt at comparing the aerial and bone-conducted components of the auditory feedback during continuous speech. We provide a measure of the informational differences between both signals, as estimated through acoustic recordings. Our results suggest that some phonemes carry different AC and BC signatures.

Several caveats are in order. First, our estimation of the BC component is dominated by soft tissue vibration, dampening the high frequencies. Second, our measure of informational difference relies on voice conversion, itself trained on limited data (∼5 min of speech). It is also defined over a limited 4 kHz bandwidth (as constrained by our estimate of BC). Last, variability in the subjects' morphological and physiological features may induce a variability of the BC component (Pollard, Tran, and Letowski 2017) that we have not investigated.

To address these issues, we are now refining our experimental and data analysis methods. We plan to improve our estimate of the BC component by adding an indirect measure of skull vibration. Besides, more participants will allow assessing across-subject variability. Last, other statistical methods may be used, e.g. the Partial Information Decomposition.

Nevertheless, these results, in particular on the differences between AC and BC feedback during consonants, suggest that unsuspected parts of the auditory feedback may affect the control of speech production. Further work should investigate the impact of bone-conducted auditory feedback on speech development and plasticity, processes in which the AC+BC sound of ones' own voice is compared to the purely AC sound of the interlocutors.

References.

Békésy, G von (May 1949). "The Structure of the Middle Ear and the Hearing of One's Own Voice by Bone Conduction". In: *The Journal of the Acoustical Society of America* 21.3, pp. 217–232.

Eeg-Olofsson, M, S Stenfelt, A Tjellström, and G Granström (2008). "Transmission of bone-conducted sound in the human skull measured by cochlear vibrations." In: *International journal of audiology* 47.12, pp. 761–769.

Hueber, T and G Bailly (2016). "Statistical Conversion of Silent Articulation into Audible Speech using Full-Covariance HMM". In: *Computer Speech and Language* 36, pp. 274–293.

Pollard, KA, PK K Tran, and T Letowski (2017). "Morphological differences affect speech transmission over bone conduction". In: *The Journal of the Acoustical Society of America* 141.2, pp. 936–944.

Pörschmann, C (2000). "Influences of bone conduction and air conduction on the sound of one's own voice". In: *Acta Acustica united with Acustica* 86, pp. 1038–1045.

Reinfeldt, S, P Östli, B Håkansson, and S Stenfelt (2010). "Hearing one's own voice during phoneme vocalization–transmission by air and bone conduction." In: *The Journal of the Acoustical Society of America* 128.2, pp. 751–762.

Stenfelt, S and RL Goode (2005). "Bone-conducted sound: physiological and clinical aspects." In: *Otology & neurotology* 26.6, pp. 1245–1261.

Toda, T and K Shikano (2005). "NAM-to-speech conversion with Gaussian mixture models". In: *InterSpeech*. Lisbon, pp. 1957–1960.

Tonndorf, J (1966). "Bone conduction. Studies in experimental animals." In: *Acta oto-laryngologica*, Suppl 213:1–133.

The production of speech modes in motor speech disorders

Marion Bourqui[1], Monica Lancheros[1], Frédéric Assal[2] & Marina Laganaro[1]

[1]Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland
[2]Department of Clinical Neurosciences, Geneva University Hospital and Faculty of Medicine, Switzerland
Marion.Bourqui@unige.ch

Introduction. Speech production is a multifaceted sensory-motor skill requiring coordination of respiratory, phonatory, and resonance systems to facilitate precise communication in diverse environments. The process involves adapting speech in response to ever-present environmental noise, necessitating variations in volume, clarity, and speed. These adaptations, referred to as "speech modes," are crucial for effective communication. The challenge arises when these modes are disrupted by motor speech disorders (MSD), particularly Apraxia of Speech (AoS) and dysarthria. AoS and dysarthria present unique challenges to speech production, stemming from distinct underlying mechanisms. AoS is characterized by laborious, non-fluent speech, syllable segregation, dysprosody, and frequent false starts (Ziegler, Aichert & Staiger, 2012), indicating impaired retrieval and assembly of phonetic plans (Varley & Whiteside, 2001). Dysarthria, on the other hand, affects speech in terms of strength, speed, range, steadiness, tone, or accuracy of movements required for speech production (Duffy, 2019). Both disorders share some clinical signs, such as phonetic distortions, reduced speech rate, and impaired prosody, but differ in their neuropathology. Few studies have directly compared these disorders, and the existing ones mostly focus on specific aspects like vowel formants (Melle & Gallego, 2012), DDK production tasks (Lancheros, Pernon & Laganaro, 2022; Ziegler, 2002), coarticulation (D'Alessandro, Pernon, Fougeron & Laganaro, 2019), and automatic discrimination (Kodrasi, Pernon, Laganaro & Boulard, 2020). Investigating speech modes within the context of MSD becomes crucial, offering valuable insights into the underlying mechanisms of these disorders. Speech modes have been investigated more extensively in dysarthria than in AoS. Hypokinetic dysarthria (HD), a subtype of dysarthria associated with Parkinson's disease, has shown difficulties in speaking loudly (Fox, Ebersbach, Ramig & Sapir, 2012) and controlling speech rate (Skodda, 2011). However, little attention has been given to AoS, possibly due to an implicit assumption that its ability to modulate speech remains unaffected. The study seeks to address this gap by exploring speech mode encoding in AoS and dysarthria. In motor speech control models literature, one model in particular proposes different processing stages governing the encoding of speech modes: the Four-level Framework (FLF, Van der Merwe, 2021). The FLF distinguishes between motor speech planning and programming, implicating the first in AoS and the latter in dysarthria. Motor speech planning encodes the place and mode of articulation, while motor speech programming defines the muscle tone, movement direction, velocity, force, range and mechanical stiffness of the muscles (Brooks, 1986). Speech modes are thus thought to be encoded at the motor speech programming processing stage, resulting in challenges for participants with dysarthria but not for those with AoS. The research design involves an experimental paradigm where participants with AoS versus HD produce speech sequences of varying syllabic complexity in standard and whispered speech. If speech modes are encoded during motor speech programming, then we expect differences in accuracy and response latencies between normal and whispered speech only in participants with suggested motor speech programming impairments, namely hypokinetic dysarthria.

Methods. The study involved 10 participants with AoS following a focal brain lesion, and 10 participants with HD associated with PD. Both participant groups were matched in severity on the BECD composite perceptive score ($p = 0.99$; Auzou & Rolland-Monnoury, 2019). The stimuli consisted of 54 disyllabic pseudo-words with varying syllabic complexity. Participants underwent a delayed production task, producing stimuli in both normal and whispered speech modes. Each trial involved a visual presentation of a pseudo-word, a delay, and a response cue for participants to produce the target stimulus. Mixed models were employed to assess performance in accuracy and initiation latencies, among the different groups with factors such as speech mode, type of syllable, group, and order of stimuli considered.

Results. no significant differences in accuracy based on speech mode was found. However, the HD group showed superior performance on illegal CCV stimuli compared to the AoS group ($p = 0.03$), but not on CV ($p = 0.88$) and legal CCV ($p = 0.75$) stimuli. In terms of initiation latencies, the AoS group exhibited longer initiation latencies in whispered speech compared to normal speech ($p < 0.001$), while the HD group exhibited faster initiation latencies in the whispered condition ($p = 0.01$). See detailed results in Figure 1.

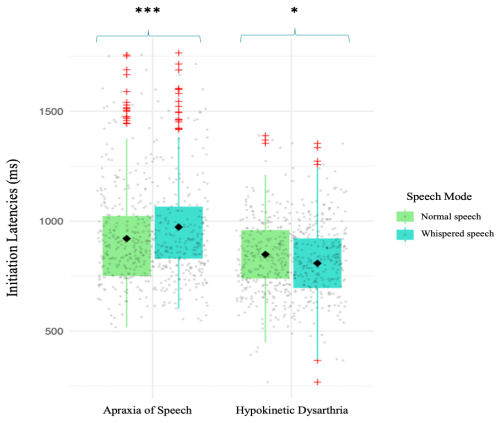


Figure 1: *Initiation latencies (ms) by group of participants, by speech mode.*

Discussion. The study explored the impact of speech mode (normal vs. whispered) on accuracy and initiation latencies in individuals with MSD, focusing AoS and HD. Notably, no significant differences in accuracy between speech modes were observed across participant groups, but the results on accuracy revealed specific challenges for the AoS group with illegal sequences. On initiation latencies however, AoS participants exhibited longer initiation latencies in whispered speech, contrary to HD participants who demonstrated faster initiation latencies in whispered speech. This unexpected contrast challenges prevailing expectations and sheds light on the distinct challenges faced by individuals with AoS and HD in non-standard speech modes. While speech modes are thought to be encoded at the motor speech programming processing stage in motor speech control models (Van der Merwe, 2021), AoS participants, which underlying deficit is supposed to be at the motor speech planning processing stage experienced challenges in whispered speech. In contrast, HD participants, which are supposed to present specific challenges in the motor speech programming processing stage, showed a surprising ease in whispered speech, contrary to previous literature that demonstrated specific challenges for this group in producing loud speech (Fox et al., 2012) or managing speech rate (Skodda, 2011). Those results for the HD group are possibly linked to the nature of extrapyramidal dysfunction in Parkinson's disease. Indeed, hypophonia is well described in HD and is marked by glottal incompetence, vocal fold bowing, and potential atrophy, stemming from rigidity in respiratory and laryngeal muscles due to the underlying extrapyramidal dysfunction in Parkinson's disease (Zarzur, Duprat, Shinzato & Eckley, 2007), accompanied by impaired scaling of vocal effort (Sapir, 2014). Taken together, the observed difficulties in whispered speech for AoS participants and the contrasting results for the HD group suggest that the interplay between motor speech planning and programming may result from a cascade effect where planning difficulties in AoS impact subsequent programming, thus causing difficulties in the production of speech modes. Those findings prompt further investigation into the mechanisms underlying different speech modes and their implications for various MSD, resulting in a reconsideration of prevailing assumptions in the field.

References

Brooks, V. B. (1986). The neural basis of motor control. Oxford University Press.

D'Alessandro, D., Pernon, M., Fougeron, C. & Laganaro, M. Anticipatory VtoV coarticulation in French in several Motor Speech Disorders. Phonetics and Phonology in Europe (PAPE 2019), Jun 2019, Lecce, Italy.

Duffy, J. R. (2019). Motor speech disorders e-book : Substrates, differential diagnosis, and management. Elsevier Health Sciences.

Fox, C., Ebersbach, G., Ramig, L., & Sapir, S. (2012). LSVT LOUD and LSVT BIG : Behavioral Treatment Programs for Speech and Body Movement in Parkinson Disease. *Parkinson's Disease*, *2012*, e391946.

Lancheros, M., Pernon, M., & Laganaro, M. (2023). Is there a continuum between speech and other oromotor tasks? Evidence from motor speech disorders. *Aphasiology, 37*(5), 715-734.

Melle, N., & Gallego, C. (2012). Differential Diagnosis between Apraxia and Dysarthria Based on Acoustic Analysis. *The Spanish Journal of Psychology*, *15*(2), 495-504.

Pernon, M., Assal, F., Kodrasi, I., & Laganaro, M. (2022). Perceptual Classification of Motor Speech Disorders : The Role of Severity, Speech Task, and Listener's Expertise. *Journal of Speech, Language, and Hearing Research*, *65*(8), 2727-2747.

Sapir, S. (2014). Multiple Factors Are Involved in the Dysarthria Associated With Parkinson's Disease: A Review With Implications for Clinical Practice and Research. *Journal of Speech, Language, and Hearing Research*, 57, 1330–1343.

Skodda, S. (2011). Aspects of speech rate and regularity in Parkinson's disease. *Journal of the Neurological Sciences*, *310*(1), 231-236.

Van Der Merwe, A. (2021). New perspectives on speech motor planning and programming in the context of the four- level model and its implications for understanding the pathophysiology underlying apraxia of speech and other motor speech disorders. *Aphasiology*, *35*(4), 397-423.

Varley, R. & Whiteside, S. (2001). What is the underlying impairment in acquired apraxia of speech. *Aphasiology, 15*, 39–49.

Zarzur, A. P., Duprat, A. C., Shinzato, G. & Eckley, C. A. (2009). Laryngeal Electromyography in Adults With Parkinson's Disease and Voice Complaints. *Laryngoscope, 117*(5), 831-834.

Ziegler, W., Aichert, I. & Staiger, A. (2012). Apraxia of Speech: Concepts and Controversies. *Journal of speech, language and hearing research, 55*, 1485-1501.

Objective measures of fatigue and sleepiness based on acoustic analysis of the temporal organization of speech

Hani Camille Yehia[1], Carla Aparecida de Vasconcelos[1], Deborah Abrante Godinho[1], Túlio Eduardo Rodrigues[2], Maurílio Nunes Vieira[3]

[1]*Graduate Program in Neuroscience, Universidade Federal de Minas Gerais, Brazil*

[2]*Institute of Physics, Universidade de São Paulo, Brazil*

[3]*Department of Electronic Engineering, Universidade Federal de Minas Gerais, Brazil*

hani@ufmg.br

Introduction. Fatigue and sleepiness are conditions that negatively affect cognitive and physical capabilities, with direct implications for performance and safety. In the particular case of aviation, fatigue is often underestimated as a factor in aeronautical events, despite its significant influence on decision-making. Psychometric scales are subjective methods for assessing fatigue and sleepiness. Objective methods are, however, important to better understand their relationship with critical events. Acoustic analysis of voice and speech has proven to be a reliable alternative for detecting a range of individual alterations and expressions, including fatigue and sleepiness, demonstrating that variations in human speech can also be robust indicators of these temporary states. The objective of this study is to investigate the applicability of objective measures of temporal organization of speech and their correlation with subjective methods as a model for detecting fatigue and sleepiness (de Vasconcelos et al. 2019).

Methods. The study is based on a retrospective cross-sectional analysis of a database containing spontaneous speech samples, sleep history data from the last 72 hours, and psychometric fatigue (Samn-Perelli) and sleepiness (Karolinska) scales of $N = 25$ aircraft pilots, native speakers of Brazilian Portuguese and neurologically healthy. The data were recorded on a day off when the pilots felt rested and on a work day after 16 hours of wakefulness. After extracting the measures of temporal organization of speech, data analysis was performed using least squares linear regression. Correlations with psychometric scales were carried out using Pearson's correlation coefficient, adopting a significance level of 5%.

Results. Measures of temporal organization of speech showed statistically significant variations when comparing rest and work days, showing a worsening of the analyzed parameters when individuals reported fatigue and sleepiness. A significant correlation was found between speech analysis measures and the Karolinska and Samn-Perelli psychometric scales. Furthermore, linear regression showed that the percentage of pauses and the speaking rate can be used to predict the level of fatigue and sleepiness reported in these scales. Among the statistically significant results obtained, it was found that there exists a high correlation ($\rho = 0.98$) between the speaking rate and the Samn-Perelli scale. This high correlation coefficient was obtained when the speaking rates measured for the $N = 25$ subjects were grouped into for bands, as shown in Figure 1, top. The 95% confidence intervals are shown in Figure 1, bottom. This result allows the level of fatigue to be inferred using an objective measure (speaking rate) instead of a subjective measure (Samn-Perelli scale).

References.

de Vasconcelos, Carla Aparecida, Maurílio Nunes Vieira, Göran Kecklund, and Hani Camille Yehia (2019). "Speech Analysis for Fatigue and Sleepiness Detection of a Pilot". In: *Aerospace Medicine and Human Performance* 90.4, pp. 415–418. DOI: doi:10.3357/AMHP.5134.2019. URL: https://www.ingentaconnect.com/content/asma/amhp/2019/00000090/00000004/art00012.

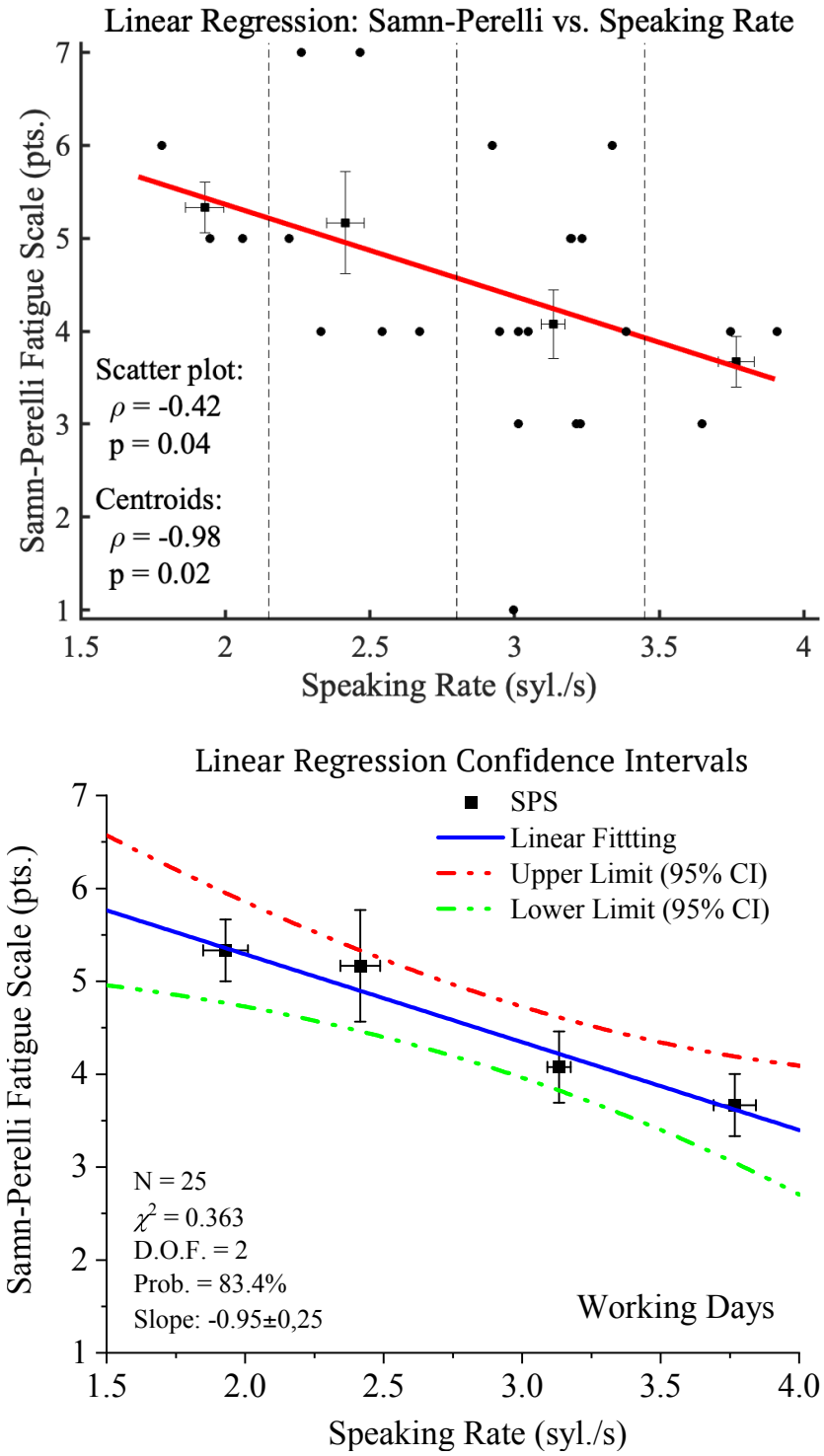


Figure 1: *Top: Linear regression showing the Samn-Perelli fatigue scale as a function of the speaking rate (red line). The results (black circles) were obtained for $N = 25$ pilots at the end of a work day after 16 hours of wakefulness. Speaking rate was measured for spontaneous speech. Subjects were asked to report events that occurred on the current day or the previous day. The recordings lasted between one and two minutes. The results were grouped into four bands indicated by vertical dashed lines. For each range, the sampled means (black squares) and the standard deviation of the sampled means (horizontal and vertical error bars) were calculated. Bottom: Linear regression (blue line) and 95% confidence interval (upper limit in red and lower limit in blue).*

The Interplay between Acoustics and Syllable Articulation Organized by Mandible Movement

Donna M. Erickson[1], Plinio A. Barbosa[2], Gustavo C. P. Silveira[2]

[1]Haskins Laboratories, New Haven, CT. USA
[2] University of Campinas, Brazil

EricksonDonna2000@gmail.com, pabarbosa.unicampbr@gmail.com,
silveira.gustavocampos@gmail.com

Introduction. Work by a number of researchers, (e.g., Erickson et al. 2012; Erickson et al. 2020; Erickson and Niebuhr in press; Erickson et al. in press; Svensson Lundmark 2023; Svensson Lundmark and Erickson 2023; Svensson Lundmark and Erickson submitted; MacNeilage 1998; MacNeilage 2008; Fujimura 2000), report that the mandible is the syllable articulator: for each syllable, the mandible opens and closes, and it is this cycle of opening and closing that defines the articulatory syllable. While the mandible is the syllable articulator, the segmental articulators are those which are crucial for making the constriction for the syllable onset and coda, during the time when the jaw is raised (closed). For example, the crucial articulator for a syllable that starts with /t/ would be the tongue tip, for a /p/, would be the lower lip, etc. Thus, the syllabic articulator and the segmental articulators are seen as separate articulatory components of a joint coordinative effort in syllable production, along the lines proposed by Fujimura (2000). As for vowel production, a pivotal articulatory work by Svensson Lundmark (2023) reports that the point in time when the crucial articulators reach peak acceleration (maximum value of the acceleration curve) is the point in time when the acoustic vowel segment starts. An acoustic study by Barbosa et al. (2016) examined velocity patterns of formant frequencies of syllable onsets to show that acoustic vowel onset coincides with maximum value of formant transition velocity. As to timing of syllable (mandible) and segmental lip opening articulation, studies by Svensson Lundmark and Erickson 2023, Svensson Lundmark and Erickson (submitted) and Erickson et al. (in press) suggest that the mandible opening for the syllable starts before the acoustic vowel while complete mandible closure occurs after the acoustic vowel. The questions we explore in this paper concern how the syllable articulator, i. e., the mandible, times its opening movements with acoustic vowel onsets as measured from spectrograms; and how this timing is affected by changes in syllable prominence, given that the jaw lowers more with increased prominence, (e.g., deJong 1995; Erickson et al. 2012; Harrington et al. 2000).

Methods. The speakers were three North American English speakers — one female (A03) and two males (A05) (A00). The utterances examined were (1) Pam said bat that fat cat at that mat, (2) Pam said BAT that fat cat at that mat, (3) Pam said bat THAT fat cat at that mat, (4) Pam said bat that FAT cat at that mat, (5) Pam said bat that fat CAT at that mat, where uppercase words indicate contrastive emphasis. Since jaw displacement varies as a function of vowel height, all syllables are closed syllables with [æ] vowels, or, in one case, [ɛ] (said). Also, notice that the target syllables are all CVC syllables. The utterances were presented to the speakers in randomized order, with five repetitions. The total number of utterances for A03 is 26, for A05 is 24 and for A00 is 31, a different number per speaker due to utterances discarded in case of problems during the acquisition of articulatory data.

Acoustic and articulatory recordings were made using 3-D EMA (Carstens AG500), courtesy of Jianwu Dang's lab at the Japanese Advanced Institute of Science and Technology, Kanazawa, Japan. One sensor was placed on the lower medial incisors (LI) to track mandible motion. Other sensors were placed on the tongue and lips, but these are not reported in this paper. Four additional sensors (upper incisors, bridge of the nose, left and right mastoid processes behind the ears) were used as references to correct for head movement. The articulatory and acoustic data were digitized at sampling rates of 100 Hz and 22.5 kHz, respectively. The occlusal plane was estimated using a biteplate with three additional sensors. In post processing, the articulatory data were rotated to the occlusal plane and corrected for head movement using the reference sensors after low-pass filtering at 20 Hz. The lowest vertical position of the LI sensor with respect to the bite plane was measured to assess how much the jaw lowered in each syllable in the utterance. In this paper, we refer only to the LI (mandible) sensor. Future work will include the other articulators in order to compare their movement characteristics with formant transitions.

Acoustic and mandible articulatory data were measured using the newly-implemented Praat algorithm (Barbosa 2023). In order to compare timing of acoustic vowel onsets (AVO) with mandible opening characteristics for each syllable, we measured with reference to AVO four extreme points in time in the articulatory data: (1) minimum value of acceleration curve associated with mandible beginning to open for vowel, (2) minimum value of velocity curve while mandible is opening, (3) maximum value of acceleration curve for when mandible was open and (4) minimum mandible position to indicate the time when mandible was maximally open for the vowel. AVO was marked manually, having the second formant transition (F2) as the reference.

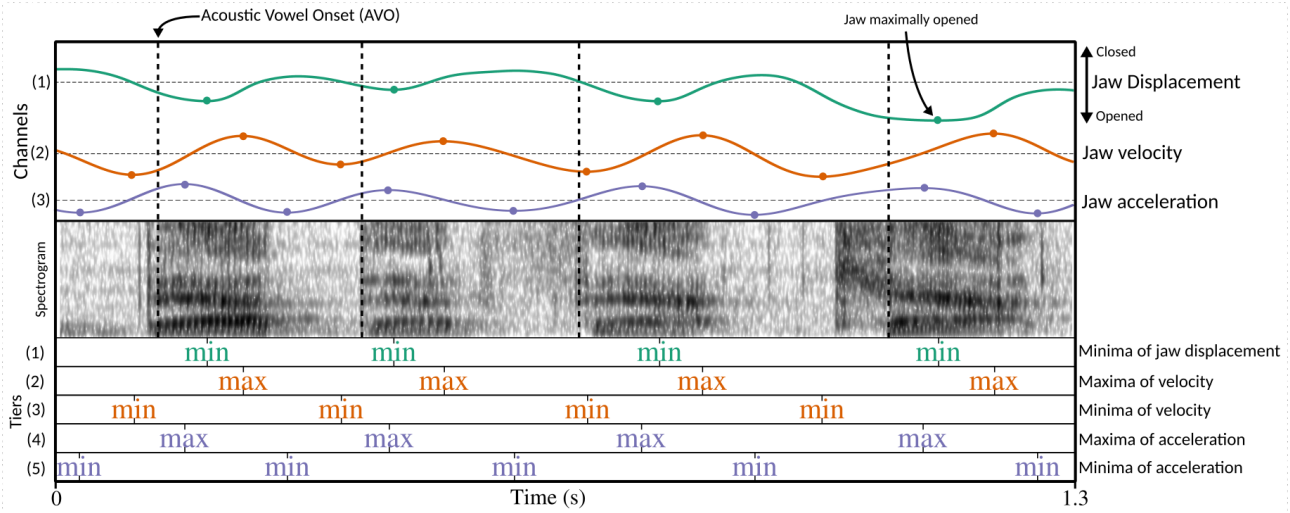


Figure 1. *Mandible vertical movement signal (Channel 1) and its first and second derivatives (Channels 2 and 3), synchronised with broadband spectrogram (0 to 5kHz), of the phrase "bat that fat CAT" by a male speaker (A00).*

Results. Initial results suggest an ordered timing of (1) minimum acceleration for the mandible beginning to open, (2) minimum velocity as the mandible is opening, (3) the acoustic vowel onset, (4) maximum acceleration as the mandible is open, (5) minimal mandible position (i.e. maximum mandible opening (displacement) for vowel), and (6) maximum velocity of the mandible as it is in the process of closing. In terms of frequency, the closest landmarks to VO were maximum acceleration (44%) and minimum velocity (40%), without a significant difference between them, followed by a significantly different maximum displacement with only 12% of times closest to VO in a proportion test (corrected α= 0.025). The median gap between maximum acceleration and VO was 30 ms, whereas that of minimum velocity and VO was 41 ms, significantly different by the Wilcoxcon test (α = 0.05). These data were obtained by a script developed by Silveira (2023). Effects of prominence, as well as voicing, manner and place of articulation of consonants, on the timing of the acceleration and velocity landmarks to AVO will also be discussed in the full paper.

Discussion. Important contributions of this paper: (1) examines the relationship between syllable (mandible) articulation and acoustic vowel onset, and (2) introduces a method to examine articulatory data. Previous methods involve Matlab or R, and may be cumbersome for many speech researchers to use. By creating a Praat script to examine articulatory data, movement position and speed, we hope it encourages more work into looking at how the articulatory signal relates to the acoustic signal.

References.

Barbosa, P. A. (2023). ConvertArticDatatoPraat. [Computer program]. https://github.com/pabarbosa/prosody-scripts

Barbosa, P. A., Madureira, S., & Camargo, Z. (2016). Scripts for the Acoustic Analysis of Speech Data. Em S. Madureira (Org.), Sonoridades/Sonorities (p. 164–174). PUC-SP.

de Jong, K. (1995). The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation. J. Acoust. Soc. Am., 97, 491–504.

Erickson, D., Suemitsu, A., Shibuya, Y., & Tiede, M. (2012). Metrical structure and production of English rhythm. Phonetica, 69, 180–190.

Erickson, D., Huang, T., & Menezes, C. (2020). Temporal organization of spoken utterances from an articulatory point of view. Proc. 10th International Conference of Speech Prosody, Tokyo, Japan, 1-5.

Erickson, D., & Niebuhr, O. (in press). Articulation of prosody and rhythm: Some possible applications to language teaching, Studies in Laboratory Phonology. Language Science Press.

Erickson, D., Svensson Lundmark, M., & Huang, T. (in press). Jaw opening patterns and their correspondence with syllable stress patterns. In Lars Meyer & Antje Strauss (Eds.) Rhythms of Speech and Language. Chapter 2.3. Cambridge University Press.

Fujimura, O. (2000). The C/D model and prosodic control of articulatory behavior. Phonetica 57, 128–138.

Harrington, J., Fletcher, J., & Beckman, M. E. (2000). Manner and place conflicts in the articulation of Australian English. In J. Broe, J.B. Pierrehumbert (eds), Papers in Laboratory Phonology, vol. 5 (p. 40–51). Cambridge: Cambridge University Press.

MacNeilage, P. F. (2008). The origin of speech. Oxford, England: Oxford University Press.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. Behav. Brain Sci., 21, 499–511.

Silveira, G. C. P. (2023). PointDistancesFromAVO. [Computer program]. https://github.com/silveira7/PointDistancesFromAVO.

Svensson Lundmark, M. (2023). Rapid movements at segment boundaries. J. Acoust. Soc. Am., 153(3), 1452–1467.

Svensson Lundmark, M., & Erickson, D. (2023). Comparing apples to oranges - asynchrony in jaw & lip articulation of syllables. In Proc. of the 20th International Congress of Phonetic Sciences, Prague, Czech Republic.

Svensson Lundmark, M., & Erickson, D. (Submitted). Segmental and syllabic articulations: a descriptive approach.

Speech Intelligibility Decreases with Degradation of Somatosensory Feedback via Topical Benzocaine Application

Elizabeth D. Casserly, Anna Barnes, Lauren Barrett

Trinity College, Hartford CT, USA

elizabeth.casserly@trincoll.edu, anna.barnes.2019@trincoll.edu, lauren.barrett.2019@trincoll.edu

Introduction. Speech sensory feedback is inherently multimodal, with information from audition, somatosensation, and proprioception combining in ongoing articulatory control. Perturbation of feedback across perceptual modalities allows us to probe the nature of the representations underlying speech production and determine how information from different sensory streams comes together in the control process. Such multisensory integration has been shown to vary across speakers in the relative strength or weighting of somatosensory versus auditory feedback, predicting the magnitude of response to perturbation in either modality (Lametti et al., 2012).

Further study into the multisensory relationship in speech feedback is warranted, but the methodologies required to perform feedback perturbation constitute a substantial barrier for widespread inquiry. The challenges are particularly large for somatosensory perturbation; effective manipulation methods have included a mechanical system linked to a custom dental apparatus that allows perturbation of the jaw with minimal acoustic effects (Lametti et al., 2012) and application of anesthetic agents performed by a physician (Larson et al., 2008). While these methods have obvious strengths, they are not accessible or feasible procedures to adopt for all speech researchers. In a series of two studies, therefore, we tested the potential for a commonly-available topical anesthetic (20% benzocaine suspension) to be used as a more accessible means of degrading somatosensory feedback in the vocal tract. In particular, we examined the impact of somatosensory degradation on perceptual intelligibility of speech, as measuring perceptual judgments from naïve listeners makes minimal assumptions regarding possible loci of change in normal speech motor control.

Experiment 1: Benzocaine degradation. Participants (N = 25) produced speech in response to 139 orthographic word prompts in two sensory conditions: baseline, with normal sensory feedback; and following application of 0.1 mL 20% benzocaine gel (Orajel) to the participants' upper and lower lips and tongue blade. Participants were shown orthographic prompts for a set of 139 English words, repeated across conditions. Lexical items were in English, selected from the Hoosier Mental Lexicon database to be rated as highly familiar to US undergraduate students (Nusbaum, Pisoni, & Davis, 1984). The stimulus set was balanced in token frequency, with 45 highly-frequent items (≥ 319 tokens/million), 45 common items (97-150 tokens/million), and 49 uncommon items (6-7 tokens/million; Nusbaum et al., 1984). Within these familiarity and frequency criteria, items were selected to maximize the use of labial articulatory gestures, containing the segments [m, p, b, f, v, ʃ, w, ɹ, i]. These segments, which incorporate labial closure, labiodental near-closure, lip rounding, and lip spreading gestures, would be most affected by the aenesthetic applied to participants' lips and therefore particularly salient to investigate in this context.

Participants' speech production across sensory conditions was recorded (Audio-Technica AT4041) in a sound-attenuating booth and used as stimuli in a "playback" perceptual study in PsychoPy (Peirce, 2007) assessing relative intelligibility judgements across the conditions. Listeners saw the orthographic representation of target words while hearing two tokens, each produced by the same talker across conditions, and completed a two-alternative force-choice task to indicate which token was "easiest to understand." Token order was randomized, and selection of items for playback was balanced across talkers and lexical frequency. We examined the rates at which listeners chose the full, undegraded baseline condition as the most intelligible, comparing it against a chance or random selection rate of 50%.

Listeners judged speech produced in the baseline as slightly more intelligible overall (M = 50.8%, SD = 3.0%), but not at rates significantly above chance ($t(32) = 1.6$, $p = .112$; see Fig. 1). Therefore, we cannot say that 0.1 mL application of benzocaine had an effect on speech intelligibility in the absence of other factors. It may be unreasonable to assume, however, that a partial disruption of somatosensory feedback would impact intelligibility in these talkers given their unimpeded access to acoustic feedback and the robust control it enables. Expt. 2 therefore used the same design to test the effects of an identical benzocaine application in the context of a previously-reported degradation of auditory feedback.

Experiment 2: Simultaneous somatosensory and auditory degradation. Manipulation of auditory speech feedback via degradation of spectral resolution has been shown to decrease talkers' intelligibility (Casserly et al., 2018). In Expt. 2, therefore, we examined the effects of benzocaine application along with simultaneous auditory spectral degradation. A new set of speakers (N = 15) completed the same speech production protocol, in baseline normal conditions and with simultaneous auditory/somatosensory degradation. Auditory degradation consisted of a real-time reduction in spectral resolution imitating the signal processing in a cochlear implant with an 8-channel noise vocoding algorithm (Casserly,

2015; Smalt, Gonzalez-Castillo, Talavage, Pisoni, & Svirsky, 2013). The PRTV hardware consisted of circumaural noise occluders (Elvex SuperSonic) with a lapel microphone fixed to the headband of the occluders above the participant's right ear, and worn beneath. The speaker's voice was detected by the lapel microphone (Williams MIC090), sent to a solid-state processor (iPod A1367) that performed the vocoding transformation using custom software, and transmitted back to noise-occluding insert earphones (Etymotic HF5) with less than 10 ms delay (Casserly, 2015). We used this particular degradation both because it allowed for direct comparison with speech produced by talkers experiencing the auditory manipulation alone – without simultaneous topical anesthetic – and because it provides a parallel sensory effect, whereby sensory information is limited in some respects (spectral resolution, labial/tongue tip somatosensation) and preserved in others (amplitude envelope, somatosensation elsewhere). Speakers in this study therefore prompted to produce the same randomized set of items as in Expt. 1, and recorded first in the baseline condition then with both the auditory spectral degradation and topical benzocaine application. Tokens of each item across conditions were then once again used in a two-alternative forced-choice playback study collecting listeners' judgments of relative intelligibility across the two conditions (listener $N = 30$).

Listeners found speech produced with the bisensory degradation significantly less intelligible than speech from the unperturbed baseline ($t(29) = 7.5$, $p < .0001$), choosing baseline tokens as "easier to understand" 59.5% of the time (SD = 6.9%). We subsequently compared this selection rate to the rates observed both with unimodal degradation of somatosensory (Expt. 1) and auditory feedback, as reported in Casserly et al. (2018). The intelligibility decrease with simultaneous benzocaine application was significantly greater than with either unimodal degradation alone (vs. somatosensory-only: indep. samples $t(61) = 6.6$, $p < .001$, Cohen's $d = 1.7$; vs. auditory only: $M_{\text{A-only}} = 55.5\%$ baseline preference, indep. samples $t(67) = 2.1$, $p = .044$, Cohen's $d = 0.5$). Figure 1 gives boxplots of the baseline-preference data from all three types of sensory degradations.

Discussion and conclusions. In these studies, we aimed to determine whether controlled application of a widely-available anesthetic used for oral pain relief might cause sufficient disruption of somatosensory feedback to impact global speech intelligibility. In Expt. 1, we did not see significant changes in naïve listeners' intelligibility judgments between speech produced with and without benzocaine application, but in the context of a simultaneous auditory degradation in Expt. 2, we did see benzocaine effects. Speakers with benzocaine *and* auditory degradation showed greater drops in intelligibility than had been observed with the same auditory manipulation alone (Casserly et al., 2018). It appears, therefore, that topical benzocaine does degrade somatosensory information to a degree relevant for speech motor control, and its effects are sufficient to impact global intelligibility when other feedback streams cannot be used to compensate. Application of benzocaine is therefore a promising new avenue for feedback perturbation research in speech to explore.

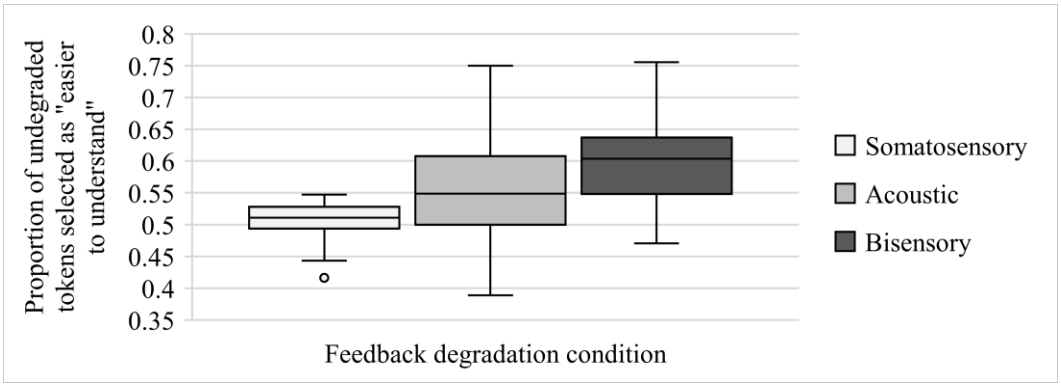


Figure 1: *Relative intelligibility judgments across the conditions of sensory degradation in Expt. 1 (left; somatosensory degradation), Expt. 2 (right; bisensory degradation), and a prior acoustic degradation (center).*

References

Casserly, E. D. (2015). Effects of real-time cochlear implant simulation on speech production. *J. Acoust. Soc. Am.*, *137*(5), 2791–2800.

Casserly, E. D., Wang, Y., Celestin, N., Talesnick, L., & Pisoni, D. B. (2018). Supra-segmental changes in speech production as a result of spectral feedback degradation: Comparison with Lombard speech. *Lang. Speech*, *61*(2), 227–245.

Lametti, D. R., Nasir, S. M., & Ostry, D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *J. Neurosci.*, *32*(27), 9351–9358.

Larson, C. R., Altman, K. W., Liu, H., & Hain, T. C. (2008). Interactions between auditory and somatosensory feedback for voice F0 control. *Exp. Brain Res.*, *187*(4), 613–621.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception Progress Report No. 10* (pp. 357–376). Bloomington, IN: Speech Research Laboratory, Indiana University.

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *J. Neurosci. Methods*, *162*(1–2), 8–13.

Smalt, C. J., Gonzalez-Castillo, J., Talavage, T. M., Pisoni, D. B., & Svirsky, M. A. (2013). Neural correlates of adaptation in freely-moving normal hearing subjects under cochlear implant acoustic simulations. *NeuroImage*, *82*, 500–509.

Manner does not modulate articulatory overlap in Spanish voiced-stop+lateral clusters

Mark Gibson[1], Stavroula Sotiropoulou[2], Adamantios Gafos[2]

[1]Universidad de Navarra
[2]Universität Potsdam
mgibson@unav.es

Introduction. Previous studies of intergestural timing in Spanish C1C2 clusters, where C1 is a voiced/voiceless stop, and C2 is a lateral /l/, have shown effects of place and voice on interconsonantal latencies (lags between the release of C1 and target of C2, or interplateau intervals, henceforth, IPI) whereby latencies in bilabial stop+lateral clusters are shorter than in velar+lateral clusters, and lags for voiced (both velar and labial)+lateral clusters are shorter than for voiceless (both velar and labial)+lateral clusters (Gibson, Sotiropoulou, Tobin & Gafos, 2019). At the same time, Spanish has a rule of approximantization such that voiced stops /b, d, g/ are produced as approximants [β, ð, ɣ] in post-vocalic contexts, even across word boundaries. Following pauses and nasal consonants they are produced as fully realized stops (see Hualde, 2014). For the present study, we examine the possible effects of manner on overlap in Spanish, where manner differences are not based on phonological contrasts (phonologically approximants are stops), but rather contextual phonetic realizations instigated by a phonological rule.

Methods. Kinematic data were collected from six native speakers of Standard Peninsular Spanish. The corpus consisted of native words that contained complex onsets involving both voiced/voiceless and labial/velar consonants in word initial position. Between 7-10 repetitions of each token were collected for each speaker. The breakdown of tokens by cluster was the following: /bl/ and /gl/ had 4 wordss each (8 total), /kl/ and /pl/ had 3 words each (6 total), yielding 620 voiced stop+lateral clusters. Words were embedded in a carrier phrase Di X, por favor, "Say X, please." The vowel /i/ (in the carrier phrase Di, "Say") preceded all target onsets. A Carstens AG501 3-dimensional electromagnetic articulograph (EMA) was used to register the kinematic movements of the different articulators. The articulatory data were recorded with a sampling rate of 250 Hz. Standard head correction procedures were performed. The reference sensors' data were filtered using a cutoff frequency of 5 Hz, while the rest of the sensors' data were filtered using a cutoff frequency of 20 Hz. Analysis of the data was carried out using MView, developed at Haskins Laboratories by Mark Tiede.

To separate the voiced stops into distinct manner categories, stops and approximants, a kernel density estimation analysis was performed to classify the phonetic category of the voiced stop, as shown in Figure 1, where two categories were fitted for the stop duration data for both /b/ (left below) and /g/ (right below). Duration was used as a metric since temporal differences between stops and approximants in Spanish correlate to constriction such that the more constriction the longer the duration (Parrell, 2011). Differences in the amplitudes of the two peaks are explained by the fact that the majority of the speakers produced approximants, as per the canonical rule for stop/approximant production in this particular context in Spanish.

In order to define a conservative threshold between the two categories, a two-component mixture model using the normalmix.EM function of the "mixtools" package (Benaglia, 2009) in R Studio was employed. After calculating the median between the two local peaks (modes), a 'find cutoff' function based on the probability of falling into a particular class was used to create two indices that define the points between the median and local peaks (both upper and lower modes) at which the probability of falling into a particular class was no better than chance (at roughly 10% above and below the median between the two local modes). The following Figure 1 shows the local peaks for the approximant and stop categories as well as the minima between the peaks, and the cutoff points for the manner categories. Tokens that fell within the cutoff thresholds, which was a parameter of the model, were discarded from the analysis:

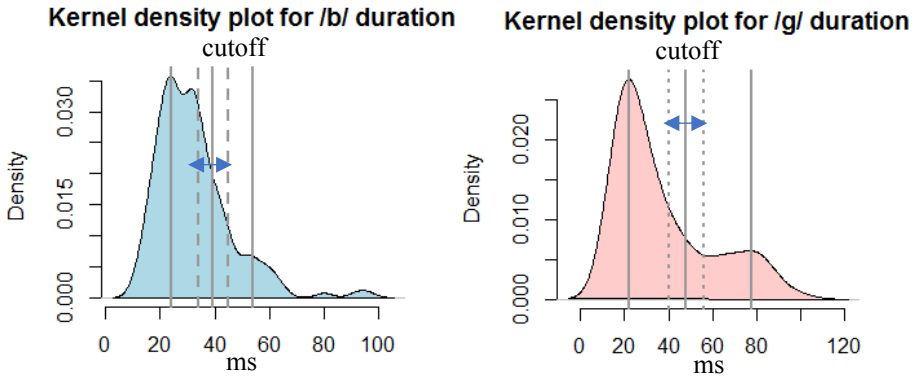


Figure 1. Kernel densities for the approximant (which corresponds to the left curve in each graph) and stop (which corresponds to the right, smaller, curve in each graph) for /b/ and /g/. Local peaks for each class (approximant and stop) are represented with a vertical solid grey line. The vertical solid grey line in the middle of these two local peaks represents the median between the two local maxima for each manner category. The dashed vertical lines represent the upper and lower cutoff thresholds for the two classes (approximants and stops). Tokens falling within the upper and lower thresholds for the two classes were eliminated from the statistical analyses.

Results. We fit a linear mixed effects model (lme4 package in RStudio, Bates et al., 2015) for manner (two levels: stops and approximants) and place of articulation (2 levels: labial and velars). Voice was eliminated for reasons of multicollinearity (all approximants were voiced). Since our previous work (Gibson et al., 2019) showed strong effects for voice on IPI, it was not necessary here to readdress the issue.

Results of our lmer models show no main effect for manner ($\chi 2$ [1, n = 620] = 0.69, p = 0.40). Mean IPI between the two manner categories varied by 6 ms (IPI for approximants = 29.1 ms, IPI for stops = 35.1 ms), as plotted in Figure 2. Interactions between place and manner were also insignificant ($\chi 2$ [2, n = 620] = 0.92, p = 0.63) in a first instance, but will be discussed more thoroughly in future studies, where we address whether variability in IPI in clusters where approximantization has taken place is perhaps attributable to less well-defined releases of the C1 stop.

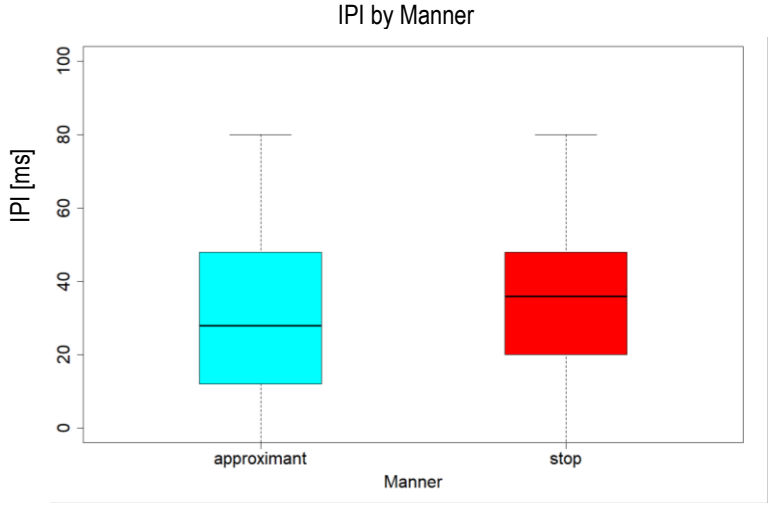


Figure 2 *Boxplots of IPI in milliseconds (y axis) for the different manner classes (x axis).*

Discussion. Our results suggest that for Spanish, manner of articulation does not modulate gestural overlap in voiced+stop lateral clusters. Our findings are of interest in light of studies that compare French and Spanish voiced stop+lateral clusters and find differences in overlap, which may be attributible to the approximantization of voiced stops in Spanish (as opposed to French). Our results suggest that manner may explain differences in overlap patterns across languages, but within the same language, where two phonetic realizations are possible for the same phonological specification (voiced stops), manner does not play a role in modulating patterns of gestural overlap.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Version 1.1.29. *Journal of Statistical Software,* 67, 1, 1−48.

Gibson, M., Sotiropoulou, S., Tobin, S., & Gafos, A. (2019). Temporal Aspects of word initial single consonants and consonants in clusters in Spanish. *Phonetica*. 1–31.

Parrell, B. (2011). Dynamical account of how /b, d, g/ differ from /p, t, k/ in Spanish: Evidence from labials. *Laboratory Phonology*, 2, 2, 423–449.

Day 2

Wednesday, May 15

08:00am		03:00pm	
08:30am	**Sophie Scott**	03:30pm	
09:00am		04:00pm	Coffee Break
09:30am	Oral Session 4	04:30pm	Poster Session 2
10:00am	**Adaptation II**	05:00pm	
10:30am	Coffee Break	05:30pm	
11:00am	Oral Session 5	06:00pm	
11:30am	**Coarticulation**	06:30pm	**Adrien Meguerditchian**
12:00am	Oral Session 6	07:00pm	
12:30am	**Phonetics/Phonology I**	07:30pm	Dinner
01:00pm		08:00pm	
01:30pm			
02:00pm	Lunch Break/free time		
02:30pm			

Oral session 4

Adaptation II

9:30– 10:30 am

	Title	Authors
9:30 - 9:50 am	Sensory errors drive speech adaptation even in the absence of overt movement	Ben Parrell (University of Wisconsin-Madison)*; Chris Naber (University of Wisconsin-Madison); Olivia Kim (Bates College); Caroline Niziolek (University of Wisconsin-Madison); Samuel McDougle (Yale University)
9:50 - 10:10 am	The Impact of Electromagnetic Articulography Sensors on the Articulatory Acoustic Vowel Space in Speakers with and without Parkinson's Disease	Thomas B Tienkamp (University of Groningen)*; Teja Rebernik (University of Groningen); Jidde Jacobi (University of Groningen); Martijn Wieling (University of Groningen); Defne Abur (University of Groningen)
10:10 - 10:30 am	The "following" paradox in studies of speech auditory-motor adaptation	Daria D'Alessandro (University of Washington)*; Nick Kitchen (Penn State University); Lisa Paroni (University of Washington); Joanne Jingwen Li (University of Washington); Elise A LeBovidge (University of Washington); Ludo Max (University of Washington)

Sensory errors drive speech adaptation even in the absence of overt movement

Benjamin Parrell[1,2], Chris Naber[2], Olivia A. Kim[3], Caroline A. Niziolek[1,2], Samuel D. McDougle[4,5]

[1]Communication Sciences and Disorders, University of Wisconsin–Madison
[2]Waisman Center, University of Wisconsin–Madison
[3]Neuroscience, Bates College
[4]Department of Psychology, Yale University
[5]Wu Tsai Institute, Yale University

bparrell@wisc.edu, cnaber@wisc.edu, okim@bates.edu, cniziolek@wisc.edu, samuel.mcdougle@yale.edu

Introduction. When we make a movement, the observed outcomes of that movement sometimes differ from our expectations. These sensory prediction errors recalibrate the brain's internal models for motor control, reflected in alterations to subsequent movements that counteract these errors (sensorimotor adaptation). Leading theories of motor learning suggest that all forms of sensorimotor adaptation are driven by learning from sensory prediction errors (Hadjiosif et al., 2021; Krakauer et al., 2019). Recent computational work has shown that adaptation to auditory perturbations in speech can similarly be driven by sensory errors through updates to internal predictive models (Kim et al., 2023). Conversely, dominant models of speech adaptation argue that adaptation results from integrating time-advanced copies of corrective feedback commands into feedforward motor programs (Guenther, 2016), which has also been suggested for reaching (Albert & Shadmehr, 2016; Kawato et al., 1987). Here, we test these alternative theories of speech adaptation by inducing planned, but not executed, speech: while the prediction error theory suggests that adaptation should only require a motor plan and a sensory error, the feedback model requires overt speech acts. Previous work in reaching has shown that limb adaptation occurs when movements are planned (generating sensory predictions) and time-aligned visual feedback containing spatial errors is given, even when the movement itself is inhibited and not executed (Kim et al., 2022). Given this result, we hypothesize that speech will similarly show adaptation driven by sensory errors even in the absence of overt speech production.

Methods. 30 speakers were prompted to speak a word and, on a subset of trials, were rapidly cued to withhold the prompted speech (**Figure 1A**). On these "no-movement" trials, just after withholding speech, speakers were exposed to playback of their own speech from the previous trial with an auditory perturbation of the first formant using Audapter (Cai et al., 2008). Similar perturbations were also applied in real time to a subset of "movement" trials to induce typical single-trial speech adaptation (Hantzsch et al., 2022). All perturbed trials occurred as the middle trial of a "triplet", preceded and followed by unperturbed movement trials. Single-trial adaptation was measured as the change in the first formant from the first to the third trial of each triplet, as measured from 50-125 ms after vowel onset to avoid 1) coarticulatory effects with the initial consonant and 2) compensatory changes that may occur starting ~150 ms into the vowel (Cai et al., 2012). Across different triplets, both upward and downward F1 perturbations were applied, and we tested for a difference in F1 adaptation between upward and downward perturbation triplets. We additionally tested for the presence of "aftereffects", lasting adaptation that persists beyond a single trial, by examining adaptation in the unperturbed trial immediately following the third trial of each triplet (this trial was either the first trial of the subsequent triplet or an unperturbed filler trial included to break up the rhythm of the perturbations). Triplets where participants produced any vocalization on the middle, perturbed trial in the no-movement condition were excluded to isolate learning in the absence of overt speech movement. Statistical analysis was conducted with linear mixed models with the lme4 package in R, including random intercepts for speaker (Bates et al., 2014; R Core Team, 2020). Statistical significance was assessed using lmerTest (Kuznetsova et al., 2017) and post-hoc comparisons were conducted with emmeans (Lenth, 2023).

Results. First, we examined the error rate on no-movement trials. All speakers failed to inhibit speech production on at least some trials (4-89 trials excluded for acoustic evidence of vocalization across participants, median = 32). This indicates that speakers likely were planning speech on no-movement trials. In terms of adaptation, our results (**Figure 1B**) indicate that speakers adapt to auditory prediction errors on both movement and no-movement trials, altering the spectral content of the spoken vowel to counteract formant perturbations (main effect of perturbation direction, $\beta = 8.7$, $t(5359.7) = 5.7$, $p < 0.0001$). However, the magnitude of adaptation is reduced on no-movement trials (interaction between direction and trial type, $\beta = 6.2$, $t(5362.3) = 5.7$, $p = 0.001$). Post-hoc tests showed a difference between trials following upward and downward perturbations in both the movement (8.63±1.51 mels, $z = 5.7$, $p < 0.0001$, $d = 0.33$) and no movement (2.45±1.21 mels, $z = 2.04$, $p = 0.044$, $d = 0.14$) triplets. Adaptation continued into the trial immediately

following each triplet (**Figure 1C**), as shown by significant aftereffects following both movement (8.47±1.69 mels, z = 5.0, $p < 0.0001$, $d = 0.30$) and non-movement (2.88±1.37 mels, z = 2.1, $p = 0.035$, $d = 0.14$) triplets.

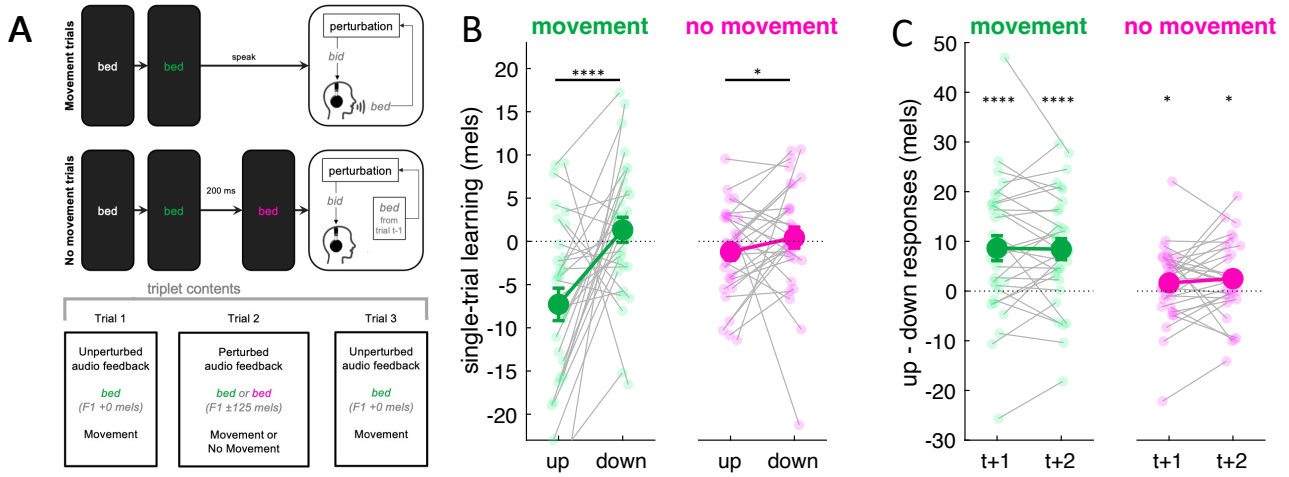


Figure 1. *A: Illustration of experiment design showing movement and no-movement trials (top) and triplet structure (bottom).* ***B****: Single-trial adaptation to upwards and downwards perturbations as measured on the third trial of each triplet.* ***C****: Differential adaptation to upwards and downwards trials on the third trial of each triplet (t+1) and on the immediately following unperturbed trial (t+2, aftereffects). * $p < 0.05$, **** $p < 0.0001$.*

Discussion. Because adaptation occurred even in the absence of any movement, these results strongly suggest that sensory prediction errors are capable of driving adaptation in speech without a contribution from online corrective movements. Given that adaptation in speech has been extensively shown to involve only implicit learning processes, the current results solidify recent observations in reaching experiments also demonstrating unconscious motor adaptation evoked by sensory errors in the absence of overt movement. Our results extend that observation to a more complex movement system, one that relies on multi-dimensional auditory feedback. Our finding that sensory errors drive adaptation in speech even in the absence of movement points to a shared computational structure across human motor systems. Moreover, the large difference observed between the upward and downward conditions in the movement trials (~9 mels) demonstrates the effectiveness of the triplet paradigm for inducing adaptation to altered auditory feedback in speech in a single trial.

References

Albert, S. T., & Shadmehr, R. (2016). The Neural Feedback Response to Error As a Teaching Signal for the Motor Learning System. Journal of Neuroscience, 36(17), 4832–4845. https://doi.org/10.1523/JNEUROSCI.0159-16.2016

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R Package Version, 1.1-12(7).

Cai, S., Beal, D. S., Ghosh, S. S., Tiede, M. K., Guenther, F. H., & Perkell, J. S. (2012). Weak responses to auditory feedback perturbation during articulation in persons who stutter: Evidence for abnormal auditory-motor transformation. PLoS One, 7(7), e41830. https://doi.org/10.1371/journal.pone.0041830

Cai, S., Boucek, M., Ghosh, S., Guenther, F. H., & Perkell, J. (2008). A System for Online Dynamic Perturbation of Formant Trajectories and Results from Perturbations of the Mandarin Triphthong /iau/. Proceedings of the 8th International Seminar on Speech Production, 65–68.

Guenther, F. H. (2016). Neural control of speech. The MIT Press.

Hadjiosif, A. M., Krakauer, J. W., & Haith, A. M. (2021). Did We Get Sensorimotor Adaptation Wrong? Implicit Adaptation as Direct Policy Updating Rather than Forward-Model-Based Learning. Journal of Neuroscience, 41(12), 2747–2761. https://doi.org/10.1523/JNEUROSCI.2125-20.2021

Hantzsch, L., Parrell, B., & Niziolek, C. A. (2022). A single exposure to altered auditory feedback causes observable sensorimotor adaptation in speech. eLife, 11, e73694. https://doi.org/10.7554/eLife.73694

Kawato, M., Furukawa, K., & Suzuki, R. (1987). A hierarchical neural-network model for control and learning of voluntary movement. Biological Cybernetics, 57(3), 169–185. https://doi.org/10.1007/BF00364149

Kim, K. S., Gaines, J. L., Parrell, B., Ramanarayanan, V., Nagarajan, S. S., & Houde, J. F. (2023). Mechanisms of sensorimotor adaptation in a hierarchical state feedback control model of speech. PLOS Computational Biology, 19(7), e1011244. https://doi.org/10.1371/journal.pcbi.1011244

Kim, O. A., Forrence, A. D., & McDougle, S. D. (2022). Motor learning without movement. Proceedings of the National Academy of Sciences, 119(30), e2204379119. https://doi.org/10.1073/pnas.2204379119

Krakauer, J. W., Hadjiosif, A. M., Xu, J., Wong, A. L., & Haith, A. M. (2019). Motor Learning. Comprehensive Physiology, 9(2), 613–663. https://doi.org/10.1002/cphy.c170043

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software, 82(13). https://doi.org/10.18637/jss.v082.i13

Lenth, R. V. (2023). emmeans: Estimated Marginal Means, aka Least-Squares Means (R package version 1.8.7) [Computer software]. https://CRAN.R-project.org/package=emmeans

R Core Team. (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org/

The Impact of Electromagnetic Articulography Sensors on the Articulatory Acoustic Vowel Space in Speakers with and without Parkinson's Disease

Thomas B. Tienkamp[1], Teja Rebernik[1], Jidde Jacobi[1], Martijn Wieling[1], Defne Abur[1]

[1]*University of Groningen*

t.b.tienkamp, t.rebernik, j.jacobi, m.b.wieling, d.abur[@rug.nl]

Introduction. Electromagnetic articulography (EMA) provides fine spatial and temporal information of articulatory movement during speech. However, the presence of the sensor coils and wires may alter the speaker's acoustic output as they might interfere with one's articulation. This raises the question to what extent the acoustic output represents the typical output of a speaker. For example, Dromey, Hunter, and Nissen (2018) found that the centre of gravity of both /s/ and /ʃ/ decreased after sensor placement and that speakers did not fully adapt acoustically 20 minutes post sensor placement. Another study found that speakers had a lower F_1 for the vowel /æ/ in the word 'black' at four minutes following sensor placement compared to directly following sensor placement (Bartholomew 2020). However, this study only assessed /æ/ and did not compare the adaptation response to the F_1 prior to sensor placement. Thus, the question remains as to what extent the presence of sensor coils affects the size of the vowel space. A better understanding of the acoustic consequences of using EMA sensors may be informative if parallel acoustic data using recordings with EMA sensors are compared to recordings without EMA sensors. Moreover, despite the increasing popularity of assessing speech motor control in clinical populations using EMA, it is currently unclear whether typical speakers and speakers with impaired vowel articulation, such as persons with Parkinson's disease (PwPD), are impacted in a comparable manner by the sensors when measured acoustically. Therefore, the objective of this study was to determine the effect of EMA sensors on sentence-level articulatory acoustic measures of speech for both typical speakers and PwPD.

Methods. This study uses data from a previous study that received ethical clearance from the institutional Medical Ethics Review Board (Jacobi 2022). We used the data from 46 individuals who gave permission for their data to be used for follow-up studies: 23 typical speakers (18 male, 5 female; mean age = 68.4 years, SD = 6.21) and 23 speakers with PD (18 male, 5 female; mean age = 69.1 years, SD = 6.98). All participants were native speakers of Dutch. PwPD participated while **ON** levodopa and had been diagnosed with Parkinson's disease between 1 and 19 years prior to their participation. All speakers read the Dutch version of the North Wind and the Sun passage before and after electromagnetic articulographic (NDI-Wave) sensors were placed on the tongue (N = 2), jaw (N = 1), and lips (N = 2). Acoustic data were assessed at three timepoints: timepoint 0 (T0), prior to sensor placement; timepoint (T1), directly after sensor placement; and timepoint (T2), at the end of the experiment, which lasted approximately an hour. Speakers were recorded with a microphone (Audio Technica AT875R) at a 22.050 Hz sample rate with a mouth-to-mic distance of approximately 20 cm. All recordings and formant traces were manually checked for abnormalities prior to data analysis. Following Whitfield and Goberman (2014), the Articulatory-Acoustic Vowel Space (AAVS) was calculated on a mel-scale based on the whole passage's continuous formant tracks of F_1 and F_2 of all voiced segments.
Linear mixed-effects models were fitted in R 4.2.0 (R Core Team 2023; Bates et al. 2015; Kuznetsova, Brockhoff, and Christensen 2017). Our model included the effect of group (PwPD vs. Typical) and time (T0, T1, T2) on the AAVS, and a random intercept per speaker. We assessed whether adding an interaction between group and time improved the fit of the model by using the *anova()* function. A significant *p*-value ($p < .05$) would indicate that the interaction improves the model. We subsequently assessed the effects of speaker sex and age in an exploratory analysis using model selection procedures, as these are known to impact vowel formants. All numerical variables were centered around the mean.

Results. The AAVS at T0 was significantly larger compared to T1 (β = 3330 mel^2, $p < .001$). There was no significant difference between T2 and T1 (β = 537 mel^2, $p = .27$). A fixed effect of group indicated that PwPD had a significantly smaller AAVS compared to typical speakers overall (β = -5705 mel^2, $p < .001$). The addition of an interaction between group and time did not improve the fit of the model ($\chi^2(2) = 1.22, p = .54$). This is visualised in Figure 1. A fixed effect

of sex indicated that males had a lower AAVS compared to females (β = -10184 mel^2, $p < .001$). Finally, a fixed effect of age indicated that AAVS decreased with speaker age (β = -187 mel^2, $p = .04$).

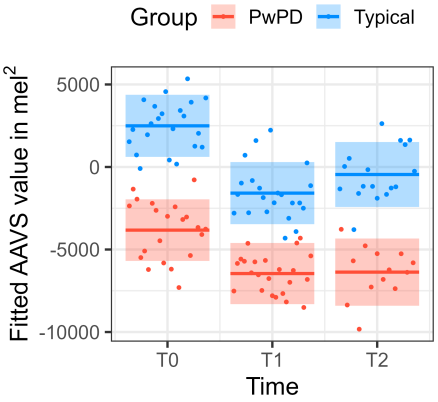


Figure 1: Model output depicting the interaction between Group and Time on the mean-centered AAVS.

Discussion. The purpose of this study was to investigate the effect of electromagnetic articulography (EMA) sensors on the articulatory acoustic vowel space (AAVS) in both typical speakers and a population that may have a reduced vowel space, namely PwPD. The results suggest that the AAVS is reduced after EMA sensor placement and does not significantly increase with habituation regardless of speaker group. On the one hand, we did not find evidence that suggests that PwPD are affected by the EMA sensors to a different extent than control speakers, which suggests that group differences were not reduced or inflated due to the placement of EMA sensors. This is relevant for both researchers and clinicians, as these results underscore the reliability of using EMA in assessing speech motor functions in PwPD. Still, PwPD did have an overall lower AAVS than typical speakers, even when accounting for sex and age differences. On the other hand, these results imply that sentence-level vowel metrics obtained from studies using both acoustic and kinematic methods might not be fully comparable to those obtained from purely acoustic designs. While Dromey, Hunter, and Nissen (2018) reported similar results for sibilants, a sound class that is actively hindered by the presence of sensors coils (i.e., through (near) sensor-palatal contact), we extend this finding by showing that EMA sensors also interfere with the vowel space as measured by the sentence-level AAVS. It remains an open question as to what extent other vowel metrics, such as the vowel space area or the vowel articulation index, are impacted by the presence of sensor coils considering these are vowel-level metrics whereas the AAVS is calculated over all voiced segments of an utterance.

References.

Bartholomew, Emily Adelaide (2020). *Kinematic and Acoustic Adaptation in Response to Electromagnetic Articulography Sensor Perturbation*. MA Thesis. Brigham Young University.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01. URL: http://www.jstatsoft.org/v67/i01/.

Dromey, Christopher, Elise Hunter, and Shawn L. Nissen (Mar. 2018). "Speech Adaptation to Kinematic Recording Sensors: Perceptual and Acoustic Findings". In: *Journal of Speech, Language, and Hearing Research* 61.3, pp. 593–603. DOI: 10.1044/2017_JSLHR-S-17-0169. URL: http://pubs.asha.org/doi/10.1044/2017_JSLHR-S-17-0169.

Jacobi, Jidde (2022). "Coordination and timing of speech gestures in Parkinson's disease". PhD thesis. University of Groningen. DOI: 10.33612/diss.238268145.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen (2017). "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13. DOI: 10.18637/jss.v082.i13. URL: http://www.jstatsoft.org/v82/i13/.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Whitfield, Jason A. and Alexander M. Goberman (Sept. 2014). "Articulatory–acoustic vowel space: Application to clear speech in individuals with Parkinson's disease". In: *Journal of Communication Disorders* 51, pp. 19–28. DOI: 10.1016/j.jcomdis.2014.06.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0021992414000562.

The "following" paradox in studies of speech auditory-motor adaptation

Daria D'Alessandro[1], Nick M. Kitchen[2], Annalisa Paroni [1], Joanne Li [1], Elise LeBovidge[1], Ludo Max[1]

[1] University of Washington, Seattle, WA, USA

[2] Penn State University, State College, PA, USA

dariada@uw.edu, nkitchen@pennstatehealth.psu.edu, lparoni@uw.edu, joanneli@uw.edu, eliseal@uw.edu, LudoMax@uw.edu

Introduction. Sensorimotor adaptation paradigms are widely used to explore the updating of feedforward movement planning based on sensory feedback experienced during prior execution of the same movements. In a typical adaptation experiment applied to speech, participants produce words or isolated vowels while their real-time auditory feedback is experimentally perturbed, for example by upward or downward shifts of the fundamental frequency (f_0) or formant frequencies (e.g., Houde & Jordan, 1998; Max et *al.* 2003, for early studies). Participants typically adapt to the feedback perturbation by making – across several trials – phonatory or articulatory adjustments such that in their own acoustic output the manipulated acoustic parameter changes in the direction opposite to the perturbation. However, most studies report also that a number of participants show not an opposing response but a *following* response. These participants change the f_0 or formants in their own acoustic output in the *same* direction as the perturbation (e.g., for f_0: Scheerer et *al.* 2016; for formants: Munhall et *al.*, 2009; MacDonald et *al.*, 2011).

Following the perturbation is a paradoxical response as it causes the experimentally induced auditory error to increase even further, rather than the error being reduced by adaptation. No information is available regarding when or why following occurs in adaptation paradigms, and the reported proportion of participants exhibiting this behavior varies considerably across studies. Interestingly, following responses also occur in studies that investigate reflexive within-trial online compensation to unpredictable f_0 perturbations rather than adaptive learning (e.g., Behroozmand et *al.*, 2012). For such f_0 compensation studies, it is known that following may be consistently exhibited by a participant or that it may vary on a trial-dependent basis with the participant demonstrating either opposing or following on different trials, and there is evidence that the variation in response direction depends on the specific state of the phonatory system at the time of perturbation onset (Franken et *al.*, 2018). At this time, it remains entirely unknown whether following in adaptation studies shows the same characteristics. Unfortunately, insight into the phenomenon is currently limited to Miller et *al.*'s (2023) recent finding that – based on data pooled across studies with different perturbations, instrumentation, and data extraction techniques – following corresponds to one tail of a unimodal distribution rather than representing a qualitatively different response in a bimodal distribution.

Here, we report on initial work aiming at better understanding following in speech adaptation paradigms by pooling data from several formant-shift perturbation studies from our own laboratory. This approach allows us to examine the individual responses of a large number of participants tested in the same laboratory setting and to assess patterns of following across well-controlled experimental conditions (e.g., sudden *vs.* gradual introduction of the perturbation) and across target vowels (e.g., front, central, back) and acoustic measures (e.g., F_1 *vs.* F_2 *vs.* aggregate measures).

Methods. For an initial analysis, the productions of 124 adult native speakers of American English who participated in seven different formant-shift adaptation studies were selected. These particular studies all involved the same overall setup but involved conditions that differed in perturbation schedule: 41 participants completed both a condition in which the perturbation was introduced suddenly and a condition in which the perturbation was introduced gradually, and 83 participants completed either only a condition with sudden perturbation or only a condition with gradual perturbation. The overall data set yielded data from 85 speakers for the sudden condition and 80 speakers for the gradual condition.

All participants produced the target words "tuck, tech, talk" (in randomized order per block of three words) repeatedly, while hearing their own auditory feedback in real-time through insert earphones. The auditory feedback signal was routed through a VoiceOne (TC Helicon) digital vocal processor under MIDI control. None of the recordings included involved masking noise or any delay in the auditory feedback signal other than the one inherent in the vocal processor itself (~10 ms). After a baseline phase with unaltered auditory feedback, a +250 cents global formant shift (i.e., applied equally to all formants) was applied either suddenly in-between two successive trials (sudden condition) or introduced gradually by ramping up from zero to maximum over many trials (gradual condition). The number of trials with the full formant shift ranged from 75 to 120 depending on the original study.

To analyze participants' response to the perturbed feedback, F_1 and F_2 from the last 5 trials of the perturbation phase were extracted at vowel midpoint and converted to semitones, expressed relative to the speakers' average F_1 and F_2 values in the baseline. F_1 and F_2 measures were considered both separately and in combination, for each target vowel separately and averaged across all target words. We then determined the mean extent of adaptation across all participants who adapted, as well as the individual response of participants who showed following for at least one formant for at least one target vowel. Since both formants were shifted upward in the auditory feedback, responses with positive values are considered following responses, while negative ones are considered adaptive responses.

Results. The number of following responses was greater in the sudden condition than in the gradual condition. Therefore, we focus here on the data for the sudden condition (85 speakers), which are shown in Figure 1. Each panel displays the average adaptive response for participants who opposed the perturbation (in grey), the individual response of participants who followed the perturbation (in color), and the percentage of followers. Across the different panels, horizontal rows show data for F_1 and F_2 separately and averaged (recall that the same formant shift perturbation was applied to both formants), and vertical columns show data for the three target vowels separately and averaged. The bottom right graph shows the total number of followers across both formants and three target vowels.

The number of identified followers varied considerably across different target vowels and formant measures. Overall, the percentage of followers was only ~8% when averaging across target vowels for either F_1 or F_2 separately, and this further decreased to only ~5% when also averaging across F_1 and F_2. Additionally, more speakers followed the perturbation when producing the word "tech" (~18%) than "talk" (~6%) or "tuck" (~11%).

Conclusions. This first analysis with 85 adult participants demonstrates clear trends in the occurrence of following responses in a speech auditory-motor adaptation paradigm with sudden onset of the perturbation. At the ISSP meeting, we will report updated results for an expanded data set with additional participants, and we will compare following data from the sudden and gradual perturbation conditions. In addition, we will present comparable data sets that enable us to assess the effect on following of methodological variables such as upward *vs.* downward formant shifts and to compare the occurrence of following responses in adults *vs.* children.

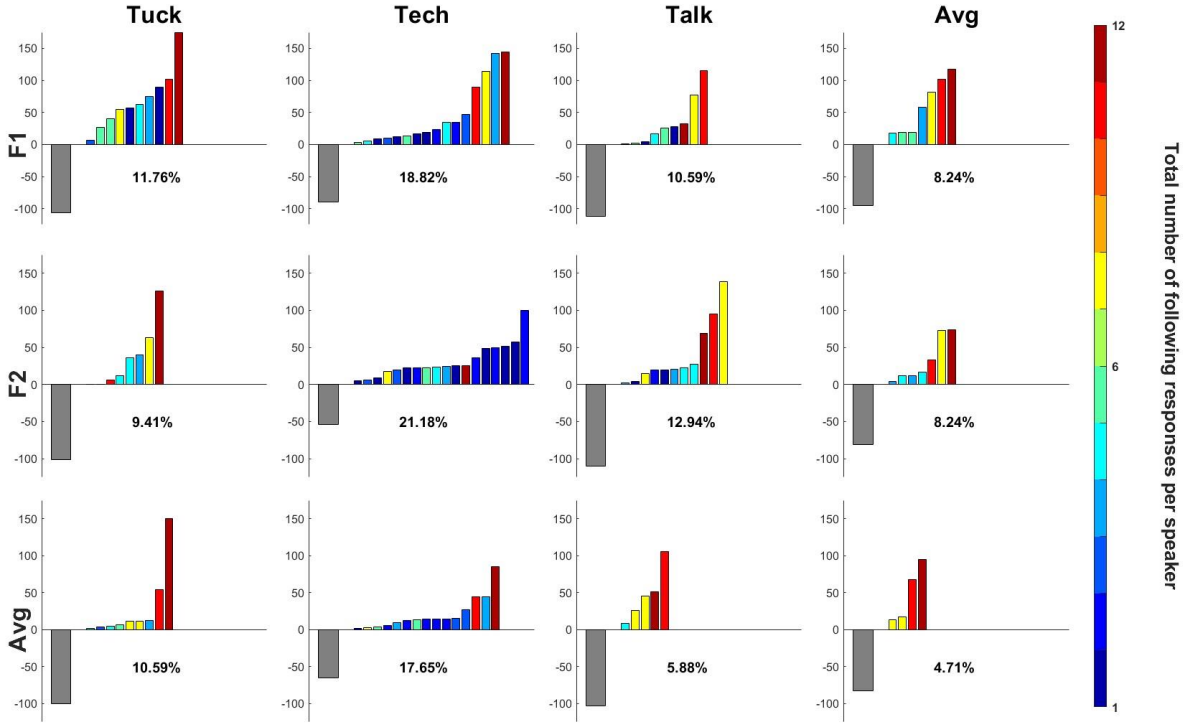


Figure 1. Average adaptation (grey) and individual following (color) by formant (rows) and target vowel (columns). Colors indicate the number of panels in which the individual showed following. Percentages indicate how many of the 85 speakers followed.

References

Behroozmand, R., Korzyukov, O., Sattler, L., & Larson, C. R. (2012). Opposing and following vocal responses to pitch-shifted auditory feedback: Evidence for different mechanisms of voice pitch control. Journal of the Acoustical Society of America, 132(4), 2468–2477.

Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., & Eisner, F. (2018). Opposing and following responses in sensorimotor speech control: Why responses go both ways. Psychonomic Bulletin & Review, 25, 1458–1467.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. Science, 279(5354), 1213-1216.

MacDonald, E. N., Purcell, D. W., & Munhall, K. G. (2011). Probing the independence of formant control using altered auditory feedback. The Journal of the Acoustical Society of America, 129(2), 955-965.

Max, L., Wallace, M. E., & Vincent, I. (2003). Sensorimotor adaptation to auditory perturbations during speech: Acoustic and kinematic experiments. In Proceedings of the 15th International Congress of Phonetic Sciences (pp. 1053-1056). Futurgraphic Barcelona, Spain.

Miller, H. E., Kearney, E., Nieto-Castañón, A., Falsini, R., Abur, D., Acosta, A., ... & Guenther, F. H. (2023). Do not cut off your tail: a mega-analysis of responses to auditory perturbation experiments. Journal of Speech, Language, and Hearing Research, 66(11), 4315-4331.

Munhall, K. G., MacDonald, E. N., Byrne, S. K., & Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. The Journal of the Acoustical Society of America, 125(1), 384-390.

Scheerer, N. E., Jacobson, D. S., & Jones, J. A. (2016). Sensorimotor learning in children and adults: Exposure to frequency-altered auditory feedback during speech production. Neuroscience, 314, 106-115.

Oral session 5

Coarticulation

11:00– 12:00 am

	Title	Authors
11:00 - 11:20 am	Children's coarticulation patterns as a window to the phonology-phonetics interface	Elina Rubertus (University of Potsdam)*; Aude noiray (LPNC- UGA)
11:20 - 11: 40 am	Coarticulation in sign language: A kinematic study on French Sign Language (LSF) using Electromagnetic Articulography (EMA)	Justine Mertz (IfL Phonetics, University of Cologne)*; Lena Pagel (IfL Phonetics - University of Cologne); Pamela Perniss (Department of Rehabilitation and Special Education, University of Cologne); Giuseppina Turco (LLF, CNRS, UMR7110, Université Paris Cité); Doris Muecke (IfL Phonetics - University of Cologne)
11:40 - 12:00 am	MRI reveals CV coarticulation is preserved in stuttering	Yijing Lu (University of Southern California)*; Louis Goldstein (University of Southern California); Shrikanth Narayanan (USC)

Children's coarticulation patterns as a window to the phonology-phonetics interface

Elina Rubertus[1], Aude Noiray[2]

[1]University of Potsdam, Germany
[2] Laboratoire de Psychologie et NeuroCognition (LPNC, UGA), France
rubertus@uni-potsdam.de, aude.noiray@univ-grenoble-alpes.fr

Introduction. One longstanding challenge for speech production theories has been to account for phonemes' discreteness on the one hand and speech continuity on the other hand. Indeed, the traditional notion of discrete and abstract phonemes is not mirrored in the articulated speech stream which neither contains clear-cut nor invariant segments but reflects dynamic articulatory movements in the vocal tract. Models of speech production have accounted for this dichotomy and the resulting effects of coarticulation in different ways. Daniloff & Hammarberg (1973) define coarticulation via phonological rules of binary feature spreading between segments. This look-ahead scanning mechanism accounts for anticipatory coarticulation but does not explain carryover effects. The authors ascribe carryover coarticulation to mechanic-inertial aspects of speech production instead. Via phonetic instead of phonological rules, the window model (Keating 1988) as well as the DIVA model (Guenther 1995; Tourville & Guenther 2011) aim to account for the graded nature of coarticulation. Both models emphasize the economy of effort for speech movements to reach phonemes' possible targets (defined by either featural specification or orosensory space, respectively). Articulatory Phonology (Browman & Goldstein 1986) in contrast, is not built on translation rules, but assumes articulatory gestures to underly both phonology and articulation. In this framework, coarticulatory effects are not interpreted as an adjustment of ideal canonical segments to their context but as overlapping invariant and intrinsically-timed gestures (Fowler, 1980).
How phonological representations are modeled into continuous speech remains debated. One way to move this research forward may be to go ontogenetically back in time and inspect earlier stages of speech production. Over the past years, the empirical work we have conducted on changes of coarticulation across childhood has provided new insights into the connection between phonology, phonetics, and articulation, and therefore informed the question of the nature of speech atoms. One relevant aspect is the development of coarticulatory strength. As Redford (2019) points out, any theory requiring computationally intensive translations from discrete, non-overlapping goals to dynamic articulatory movements, predicts a slow increase of coarticulatory strength across childhood. Another point is the dichotomy of underlying mechanisms for anticipatory on the one hand and carryover coarticulation on the other hand: If anticipatory behavior is planned while carryover coarticulation results from motoric constraints and muscle inertia, their evolution may differ greatly across childhood, while a common mechanism like coproduction would imply parallel developments.

Methods. To address those questions, we recorded 75 German native speakers in 5 different age groups (3y, 4y, 5y, 7y, and adults) within SOLLAR (Noiray et al. 2020), a child-friendly recording platform combining ultrasound tongue imaging, acoustic, and video data. In an acoustic repetition task, participants produced C_1VC_2ə pseudowords (C = /b/, /d/, /g/, V = /i/, /y/, /u/, /a/, /e/, /o/, $C_1 \neq C_2$) preceded by the article /aɪnə/. Using generalized additive mixed modeling (GAMM; Wood 2017), we investigated vowel-induced horizontal displacement of the tongue dorsum's highest point and its interaction with age and consonant identity at four vowel-preceding (Noiray, Wieling, Abakarova, Rubertus, & Tiede, 2019) and four vowel-following (Rubertus & Noiray, 2020) time points.
For a comparison of read versus repeated speech, an additional 32 7- to 9-year-old children and 16 adults were recorded within the same setup. We analyzed anticipatory coarticulation in this data set using 23 vowel-preceding time points and comparing the resulting movement trajectory of the tongue dorsum over time between stimuli with front vowel /i/ and those with back vowel /u/ again using GAMMs (Rubertus, Popescu, & Noiray, n.d.).

Results. In both the anticipatory and the carryover directions, coarticulatory strength substantially decreased with age. The youngest child cohort exhibited strongest vocalic coarticulation while adults' coarticulation was weakest. Our data yielded further important results:

- Children's utterances displayed discontinuous effects of carryover coarticulation: While there was no, or very limited vocalic information present within the temporal domain of the consonant /d/, tongue positions during the following schwa were vowel-dependent again.
- Intervocalic consonants did not affect anticipatory V-to-V coarticulation degree in children but in adults.

- Beginning readers showed limited coarticulatory extent when reading aloud compared to repeating stimuli, while proficient (adult) readers did not differ in coarticulation extent between these modalities.

Discussion. The developmental decrease of coarticulatory degree we repeatedly found in our empirical studies along with others across languages (e.g., Zharkova, Hewlett, & Hardcastle, 2011) is problematic for speech production models arguing for pre-planning and complex translation mechanisms from the underlying segments to their implemented form in the vocal tract. The coproduction framework (Fowler, 1980) provides a plausible explanation for the developmental change: It ascribes context-effects to low-level interactions of temporally overlapping coordinative constraints during the articulatory implementation of linguistic segments. This conceptualization of coarticulation as gestural coproduction is also supported by the parallel developments in anticipatory and carryover coarticulation highlighted in our studies. Moreover, the observed discontinuous vocalic effects as well as the lack of impact from intervocalic consonants on children's V-to-V coarticulation degree provide evidence for invariantly broad vocalic activation with temporally very limited consonantal clamps of the tongue. In that perspective, the developmental decrease of coarticulatory degree may be envisioned as a compression of vocalic activation curves leading to a gradual limitation of overlap between the vowel's and surrounding gestures (cf. Nittrouer 1993, Figure 1).

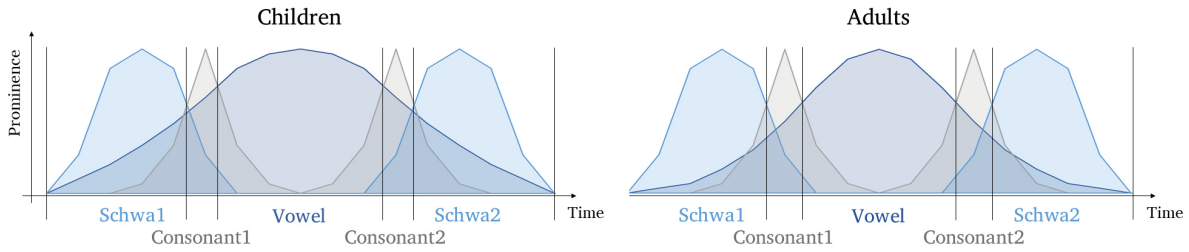


Figure 1: *Segments' hypothesized prominence over time in utterances of the form əCVCə.*

Literacy acquisition may be one factor stimulating a decrease in the width of vocalic activation curves. Indeed, exposition to alphabetic orthographies may not only raise awareness of phoneme-sized units but may also contribute to reducing the relative prominence of stressed vowels. We intend to pursue this work in future empirical investigations.
As promoted by Vihman and Croft (2007), Redford (2019), and others, our work highlights that investigations of coarticulatory changes across childhood are not only essential for our understanding of spoken language development, but that the developing system provides important insights into the phonology-phonetics interface.

References

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*(1986), 219–252.

Daniloff, R. G., & Hammarberg, R. E. (1973). On defining coarticulation. *Journal of Phonetics*, *1*(3), 239–248.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*(1), 113–133.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*(3), 594.

Keating, P. A. (1988). The window model of coarticulation: articulatory evidence. *UCLA Working Papers in Phonetics*, *69*, 3–29.

Nittrouer, S. (1993). The emergence of mature gestural patterns is not uniform: Evidence from an acoustic study. *Journal of Speech, Language, and Hearing Research*, *36*(5), 959–972.

Noiray, A., Ries, J., Tiede, M., Rubertus, E., Laporte, C., & Ménard, L. (2020). Recording and analyzing kinematic data in children and adults with SOLLAR: Sonographic & Optical Linguo-Labial Articulation Recording system. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *11*(1), 14.

Noiray, A., Wieling, M., Abakarova, D., Rubertus, E., & Tiede, M. (2019). Back from the future: non-linear anticipation in adults' and children's speech. *Journal of Speech, Language, and Hearing Research*, *62*(8S), 3033–3054.

Redford, M. A. (2019). Speech production from a developmental perspective. *Journal of Speech, Language, and Hearing Research*, *62*(8S), 2946–2962.

Rubertus, E., & Noiray, A. (2020). Vocalic activation width decreases across childhood: Evidence from carryover coarticulation. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *11*(1), 7.

Rubertus, E., Popescu, A., & Noiray, A. (n.d.). The protracted development of phonemic blending fluency is reflected in coarticulatory patterns: Evidence from beginning and proficient readers. *In Preparation*.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952–981.

Vihman, M. M., & Croft, W. (2007). *Phonological development: Toward a "radical" templatic phonology*.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Zharkova, N., Hewlett, N., & Hardcastle, W. J. (2011). Coarticulation as an indicator of speech motor control development in children: An ultrasound study. *Motor Control*, *15*(1), 118–140.

Coarticulation in sign language: A kinematic study on French Sign Language (LSF) using Electromagnetic Articulography (EMA)

Justine Mertz[1,2], Lena Pagel[1], Pamela Perniss[3], Giuseppina Turco[2], Doris Mücke[1]

[1]*IfL Phonetics, University of Cologne*

[2]*Laboratoire de Linguistique Formelle, CNRS, UMR 7110, Université Paris Cité*

[3]*Department of Rehabilitation and Special Education, University of Cologne*

jmertz1@uni-koeln.de, lena.pagel@uni-koeln.de, pperniss@uni-koeln.de

giuseppina.turco@cnrs.fr, doris.muecke@uni-koeln.de

Introduction. Coarticulation is a crucial aspect of communication during interactions. When unconstrained by perceptual demands, the speech motor system tends to minimize the physical costs of the speech system leading to a higher overlap of articulatory movement patterns (Lindblom 1990). Anticipatory coarticulation in spoken language underlines speakers' adaptation to the complex demands of the communication process by reducing or increasing articulatory effort, and this behavior supports listeners' predictions of forthcoming information (Liberman and Mattingly 1985). So far, the role of coarticulatory strategies in sign language (SL) is unclear. Previous research demonstrated anticipatory movements in handshape and/or location in American Sign Language (ASL; e.g., Cheek 2001; Gurbuz et al. 2021; Mauk, Lindblom, and Meier 2008; Tyrone and Mauk 2010), using various methodologies such as motion capture, Radio Frequency sensing (RF-sensing) or manual-based video annotation. In spoken language, kinematics of coarticulation are observed through techniques like 3D-Electromagnetic Articulography (EMA). We present the first study on *coarticulation* in a signer's production in French Sign Language (LSF) using EMA to quantify his behavior for balancing transmission accuracy and resource costs. In this novel approach, we aim to demonstrate the methodological adjustments needed to conduct a more extensive study with EMA in the study of SLs, focusing on the anticipatory coarticulation patterns observed in one signer.

Methods. One native deaf signer of LSF was recorded with EMA while facing a computer monitor displaying the stimuli. To track the articulators' movements, EMA sensors were placed on the head, torso, arms and fingers. The task consisted in the production of phonological pairs of signs (reported here as X1 and X2) composed of '1'- and/or '3'-handshape varying in location (forehead, mouth, neutral space): '1'-handshape corresponds to the index finger extended (GERMAN, ORDER, HAVE-TO) and '3'-handshape to thumb, index and middle fingers extended (ROOSTER, BAR, APARTMENT). To capture finger extension/closing, sign combinations included target pairs with X1 having the '1'- and X2 the '3'-handshape, resulting in a pair '1-3', or vice versa, resulting in a pair '3-1' (total of 18 pairs). Control pairs included '1-1' and '3-3' handshapes (limited to 4 pairs). Each pair was produced three times (total of 66 trials). Kinematic recordings were performed using 3D EMA (AG501) and a time-synchronized video set-up. We used ELAN (Crasborn and Sloetjes 2008) for video annotation and signal alignment of EMA transformed data in each trial. The kinematic offset of X1 and onset of X2 were annotated for the analysis based on articulatory landmarks, which were defined for each sign (e.g., position of the wrist on the y-axis for HAVE-TO as X1). Then, the 3D euclidean distance between the thumb and the pinkie finger was measured to capture the extension of fingers in '1-3' combinations (= increase of distance), and closing of fingers in '3-1' combination (= decrease). Onset and target achievement of the extension/closing, its peak velocity and peak acceleration were then detected automatically. The data were analyzed in the framework of dynamical systems (Task Dynamics/Articulatory Phonology, Browman and Goldstein 1992; Kelso 1995; Gafos and Benus 2006; Mücke, Hermes, and Cho 2017) that allows for the direct mapping of phonological information (low-dimensional description) onto continuous phonetic cues (high-dimensional description) in SL. This framework allows for quantification of coarticulatory patterns in SL, e.g., with respect to different speaking styles or communicative demands. By the time of the conference, statistical analyses will be carried out using linear mixed models.

Results. The articulatory data show evidence of coarticulation. We see anticipatory movements of handshape change

before the end of X1 in several trials, including various signs in both '1-3' and '3-1' combinations. An example is provided in **Fig. 1** below. In signs with repetitive movement (i.e., ROOSTER, GERMAN, APARTMENT, BAR), the kinematic data allows to detect the full repetition of the movement that was not always visible on video data, as well as partial and full truncation of the movement. A detailed description of the productions will be presented at the conference.

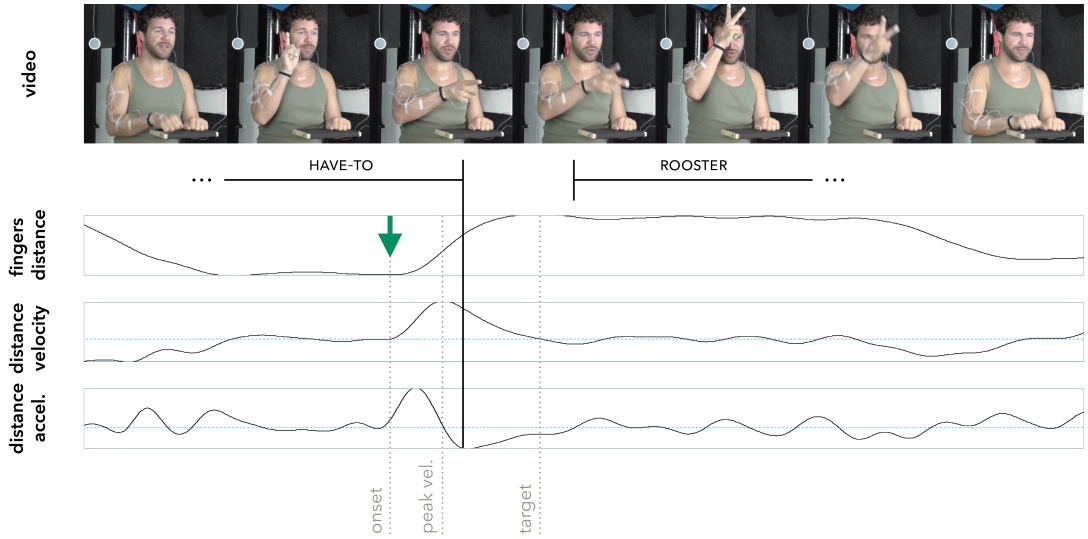


Figure 1: *Example of coarticulation in the sequence* HAVE-TO - ROOSTER *('1-3'). 3D euclidean distance between the thumb and the pinkie finger, velocity and acceleration of* ***extension*** *show that the onset starts before the end of* HAVE-TO.

Discussion. The use of 3D EMA in SL research has proven to be highly effective, enabling precise kinematic measurements and the analysis within a dynamical framework. Our preliminary exploration of EMA set-ups facilitated the development of a meticulously controlled experimental design, mitigating technical challenges associated with the 3D extent of signing space in front of and on the body due to the electromagnetic field's limitations (e.g., restrictions on sensor distance and height). In comparison to RF-sensing, EMA can capture gradient changes in handshape and non-manual components, offering a more cost-effective alternative to motion capture. This advancement holds significant potential for the development of dynamical descriptions of SL, providing a valuable tool for studying multi-modality domains such as co-speech gestures and non-manual components in SL lexicons. This methodological approach extends its utility to bilingual bimodal speech, encompassing visual cues for communication purposes and even the analysis of mouthing. We look forward to presenting a **comprehensive overview of our methodological investigation** and our experimental results during the conference, outlining the adaptations made for kinematic recordings with EMA.

References.

Browman, C. P. and L. Goldstein (1992). "Articulatory phonology: An overview". In: *Phonetica* 49.3-4, pp. 155–180.

Cheek, A. (2001). "Synchronic handshape variation in ASL: evidence of coarticulation". In: *NELS* 31.1, p. 9.

Crasborn, O. and H. Sloetjes (2008). "Enhanced ELAN functionality for sign language corpora". In: *LREC 2008*, pp. 39–43.

Gafos, A. I. and S. Benus (2006). "Dynamics of Phonological Cognition". In: *Cogn. Sci.* 30.5, pp. 905–943.

Gurbuz, S. Z., A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. S. Crawford, M. M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, and R. Mdrafi (2021). "American Sign Language Recognition Using RF Sensing". In: *IEEE Sensors Journal* 21.3, pp. 3763–3775.

Kelso, JA S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. MIT press.

Liberman, A. M. and I. G. Mattingly (1985). "The motor theory of speech perception revised". In: *Cognition* 21.1, pp. 1–36.

Lindblom, B. (1990). "Explaining Phonetic Variation: A Sketch of the H&H Theory". In: *Speech Production and Speech Modelling*. Ed. by William J. Hardcastle and Alain Marchal. NATO ASI Series. Dordrecht: Springer Netherlands, pp. 403–439.

Mauk, C. E., B. Lindblom, and R. P. Meier (2008). "Undershoot of ASL locations in fast signing". In: *Signs of the time* 8, pp. 3–24. (Visited on 12/14/2023).

Mücke, D., A. Hermes, and T. Cho (2017). "Mechanisms of regulation in speech: Linguistic structure and physical control system". In: *J. Phon.* Mechanisms of regulation in speech 64, pp. 1–7.

Tyrone, M. E. and C. E. Mauk (2010). "Sign lowering and phonetic reduction in American Sign Language". In: *J. Phon.* 38.2, pp. 317–328.

MRI reveals CV coarticulation is preserved in stuttering

Yijing Lu[1], Louis Goldstein[1], Shrikanth Narayanan[1,2]

[1]Department of Linguistics, University of Southern California
[2]Ming Hsieh Department of Electrical and Computer Engineering
yijinglu@usc.edu, louisgol@usc.edu, shri@ee.usc.edu

Introduction. Stuttering is fluency disorder described as featuring three distinct types of dysfluencies in speech, typically associated with syllable-initial consonants: (1) repetitions of one or more speech sounds (e.g., "m-m-m-mom", "be-be-be-because"); (2) prolongations of speech sounds (e.g., "wwwhere", "llllook"); (3) tense pauses prior to producing speech sounds, known as blocks (e.g., "–dad", "ea–ten"). These traditional hallmark characteristics of stuttering are defined from an auditory-perceptual point of view, while the articulatory behaviors that give rise to them remain understudied and poorly understood. Consequently, theories of stuttering often rely on subjective inferences regarding the behaviors underlying auditory manifestations, potentially leading to misconceptions about the actual breakdowns in stuttering. One notable example is that several theories of stuttering (Guenther, 2016; Howell, 2004; Postma & Kolk, 1993) propose that stuttering stems from disruptions in initiating or planning the next phoneme, assuming that the next phoneme is not present during stuttering moments. However, the failure of the annotator to hear the next phoneme does not necessarily mean its articulatory gesture is absent. Numerous articulatory studies (e.g., Goldstein et al., 2007; Hagedorn et al., 2017) have shown that supra-laryngeal articulatory gestures can take place in the vocal tract without generating auditory or acoustic consequences. Therefore, although it sounds like the next phoneme starts after the dysfluency ends, it is possible that the articulatory gesture associated with that phoneme has already started during the dysfluency. To test this hypothesis, the current study examines the tongue gesture during syllable-initial labial dysfluencies to determine if the tongue gesture for the syllable nucleus (vowel) is already being produced during the syllable-initial consonant dysfluencies.

Methods. A 0.55T MRI system (Siemens Aera XQ) was used to acquire image data from 8 adults who stutter (7 male, 1 female) during a passage reading task (three passages, Rainbow, Grandfather, Northwind and Sun, repeated twice), with a noise cancelling microphone (OptoAcoustics FOMRI-III) used to collect audio data concurrently with the RT-MRI data acquisition. A 13-interleaf spiral-out balanced steady-state free precession (bSSFP) sequence was used. Imaging parameters were: repetition time = 5.03 ms, echo time = 0.7694 ms, field-of view (FOV) = 240 × 240 mm^2, slice thickness = 6 mm, spatial resolution = 2.3 × 2.3 mm^2, flip angle = 35 degree. Images were reconstructed with a temporal resolution of 10.06 ms per frame (99.4 frames per second).

Speech data analyzed in this study are syllables with dysfluencies on the initial labial consonant and the syllable nucleus being either a high front vowel (which requires a palatal constriction) or a low back vowel (which requires a pharyngeal constriction). These combinations of consonants and vowels are chosen because the articulatory gestures they require minimally interfere with each other. Region-of-interest (ROI) analysis (Lammert *et al.* 2010; Blaylock 2021) is used to capture the change of labial constriction degree, palatal constriction degree, and pharyngeal constriction degree over time, indicated by the time series of mean pixel intensity within the corresponding ROI (Figure. 1a). Pixel intensity time series are smoothed using a wavelet transform-based method to remove small fluctuations caused by MRI image noise instead of articulator movement. If the vowel gesture has already been initiated during dysfluencies, we expect to see the palatal and pharyngeal constriction degrees differ as a function of the phonetic category of the following vowel.

Results. The lip constriction time series during labial dysfluencies showed two basic patterns: 1. lip fixation, resulting in overly sustained labial constriction; 2. oscillatory movement of lips, resulting in consecutive lip closing and opening (Figure 1b). This is consistent with previous findings in Zimmermann (1980) and Lu et al. (2022). The observation window in which the vowel gestures are examined is defined as the time period from the onset of the constriction plateau to its offset in the case of fixation, and from the achievement of the first constriction to the offset of the last constriction in the case of oscillation. Given that fixation and oscillation sometimes co-occur in the same instance of dysfluency, in those mixed cases, starting and ending points of the observation window are taken according to the local patterns. 20% velocity thresholding is used to determine the gestural kinematic landmarks. Means values of palatal and pharyngeal constriction degree measurements during the observation window of each disfluency were calculated (boxplots shown in Figure 1c).

For all the participants, the palatal and pharyngeal constriction degrees during the observation window showed a systematic differentiation across the two vowel conditions: the palatal constriction degree is higher in the high front vowel condition than in the low back vowel condition; the pharyngeal constriction degree is higher in the low back vowel condition than in the high front vowel condition (Figure 1c). Sign tests that compare the mean palatal and pharyngeal constriction degrees between the two vowel conditions across eight participants reveal a statistically significant difference

(p = 0.0078). The difference is still significant (p = 0.0078) when the observation window is reduced to the first half of the original window.

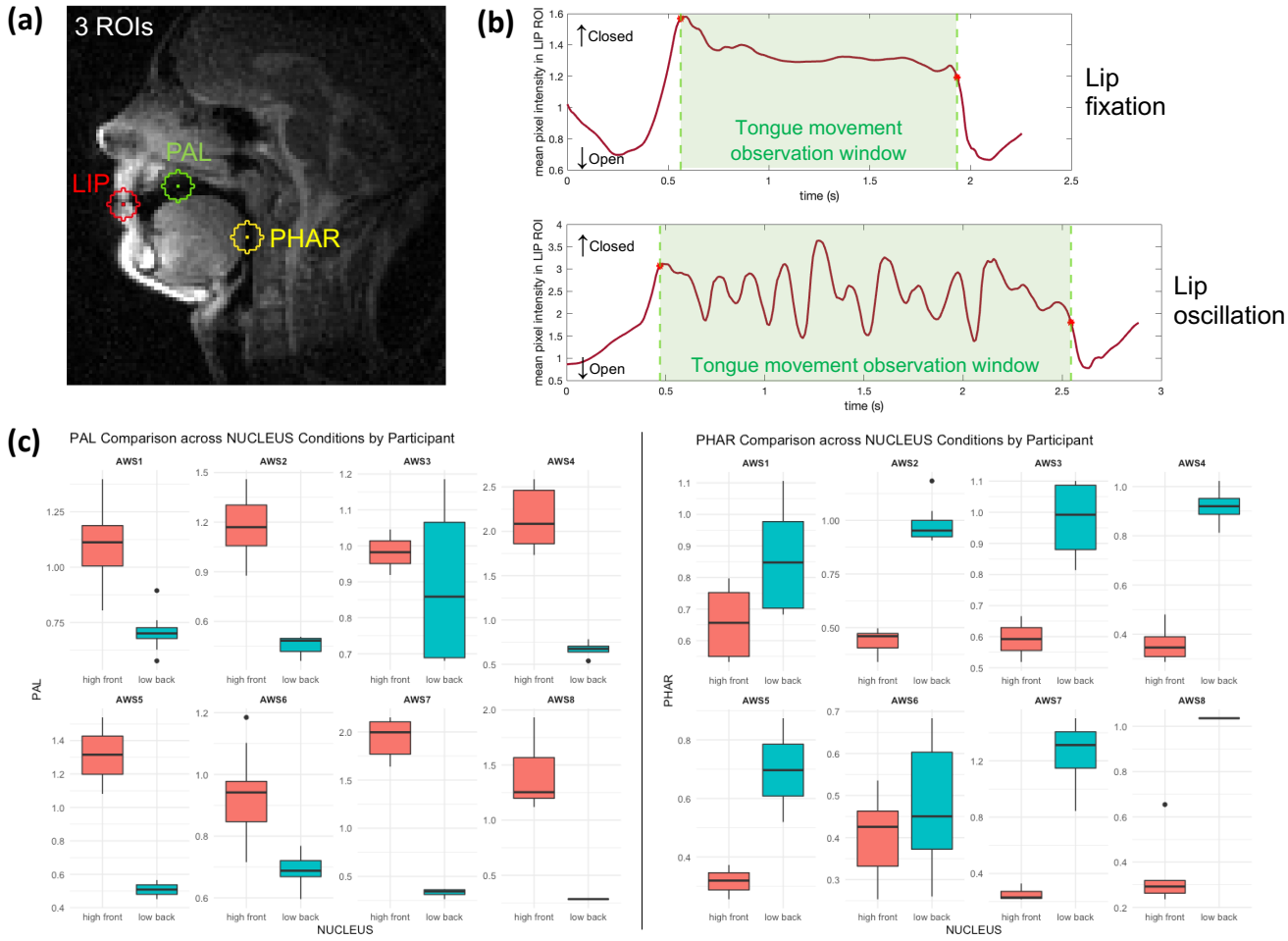


Figure 1: *(a) Three ROIs: lip (red), palatal (green), pharyngeal (yellow); (b) Two patterns of dysfluent gestures; (c) palatal and pharyngeal constriction degree measurements during the dysfluencies.*

Discussion. Contrary to the auditory perception that the syllable nucleus has not started during the onset consonant dysfluencies, this study provides evidence showing that the articulatory gesture for the nucleus vowel has already been initiated in the vocal tract way before the end of the dysfluency, in fact within the first 50% of the dysfluency. This result challenges the hypothesis that the stuttering dysfluencies stem from the problems with initiating or planning the upcoming phoneme. The coarticulation between the consonant and vowel gestures is preserved during stuttering, despite that the consonant gesture shows atypical kinematic patterns.

References

Blaylock, R. (2021). VocalTract ROI Toolbox. Available online at https://github.com/reedblaylock/VocalTract-ROI-Toolbox.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition*, 103(3), 386-412.

Guenther, F. H. (2016). *Neural control of speech*. Cambridge, MA: MIT Press.

Hagedorn, C., Proctor, M., Goldstein, L., Wilson, S. M., Miller, B., Gorno-Tempini, M. L., & Narayanan, S. S. (2017). Characterizing articulation in apraxic speech using real-time magnetic resonance imaging. *Journal of Speech, Language, and Hearing Research*, 60(4), 877-891.

Howell, P. (2004). Assessment of some contemporary theories of stuttering that apply to spontaneous speech. *Contemporary Issues in Communication Sciences and Disorders*, 31, 123–141.

Lammert, A. C., Proctor, M. I., & Narayanan, S. S. (2010). Data-driven analysis of real-time vocal tract MRI using correlated image regions. *Proceedings of INTERSPEECH 2010*.

Lu, Y., Wiltshire, C. E., Watkins, K. E., Chiew, M., & Goldstein, L. (2022). Characteristics of articulatory gestures in stuttered speech: A case study using real-time magnetic resonance imaging. *Journal of Communication Disorders*, 97, 106213.

Postma, A., & Kolk, H. (1993). The covert repair hypothesis: Prearticulatory repair processes in normal and stuttered disfluencies. *Journal of Speech, Language, and Hearing Research*, 36(3), 472-487.

Zimmermann, G. (1980b). Articulatory behaviors associated with stuttering. *Journal of Speech Language and Hearing Research*, 23(1), 108-121.

Oral session 6

Phonetics/Phonology I

12:00 am– 01:00 pm

	Title	Authors
12:00 - 12:20 am	Tonal-segmental interaction in diphthong realization in Standard Mandarin	Chenyu Li (Université Paris Cité, LLF, CNRS)*; Jalal Al-Tamimi (Université Paris Cité)
12:20 - 12:40 am	The interaction between phonetics and phonology when processing the acoustic signal: evidence from labial coarticulation in English and French	Phil J Howson (Ludwig-Maximilians-Universität München)*; Marianne Pouplier (LMU); Francesco Rodriquez (Ludwig-Maximilians-Universität München); Eva Reinisch (Austrian Academy of Sciences); Justin Lo (University College London); Chris Carignan (University College London); Bronwen Evans (University College London)
12:40 am - 1:00 pm	Production Allophones of North American English Liquids	Mark Tiede (Yale University)*; Suzanne Boyce (University of Cincinnati); Michael Stern (Yale University); Teja Rebernik (University of Groningen); Martijn Wieling (University of Groningen)

Tonal-segmental interaction in diphthong realization in Standard Mandarin

Chenyu Li[1], Jalal Al-Tamimi[1]

[1] Université Paris Cité, CNRS, Laboratoire de Linguistique Formelle, F-75013 Paris, France
Chenyu.li@etu.u-paris.fr, jalal.al-tamimi@u-paris.fr

Introduction. For the research of speech sound, the phonological approaches often consider the segmental features, as the height or backness of the vowels, and the suprasegmental information, as the stress or tones, separately. However, with accumulating evidence of intrinsic fundamental frequency (*f*0) of the vowels and the tonal effect on vocalic realizations of simple vowels, the interaction between tones (contours of *f*0) and vocalic segments has been extensively confirmed in terms of monophthongs. The intrinsic *f*0, suggesting a general tendency for high vowels to have higher *f*0 and vice-versa (Chen *et al*. 2021), has been documented as a universal phenomenon across language, either tonal (Wang 2007) or non-tonal (Whalen & Levitt 1995). Regarding the effect of *f*0 on vowel realization, which is referred to as the tonal-segmental interaction within tonal languages, different studies have evaluated it from an articulatory (e.g., Shaw *et al.* 2016) and an acoustic (e.g., Wang 2007; Erickson *et al*. 2004) viewpoint. In general, a higher *f*0 will cause the vowel, especially the /a/ to be realized as higher and more front. The tonal effect on high vowels such as /u/ and /i/, often has an "inverse effect", which is possibly due to the different mechanism of the larynx-vocal tract linkage (Chen *et al.* 2021) or due to the phonological control by the speaker in high vowel range (Shaw *et al.* 2016).
Concerning the tonal-segmental interaction in a dynamic situation, e.g., diphthong, Li *et al.* (2023) and Li & Al-Tamimi (under review) demonstrate that in Standard Mandarin, in line with previous studies, the correlation between tone (*f*0) and vowel realization also exists significantly in the diphthong /ai/. A high tone (tone1 in Standard Mandarin) results in a higher and more front /ai/ and a low tone (tone3) results in a lower and more back realization; a rising tone (tone2) strengthens the dynamic features of /ai/ while a falling tone (tone4) weakens such features and leads monophthongization. The results also suggest that the relation between *f*0 and vowel realization, found in simple vowel /a/, is equally applicable to the second target in /ai/ which is /i/: *f*0 is negatively correlated with F1 and positively with F2.
The present study aims to explore whether the tonal influence on the diphthong /ai/ realization in Standard Mandarin found in Li *et al.* (2023) and Li & Al-Tamimi (under review) also exists in the /au/ case. We hypothesize that for /au/, the vowel realizations will be affected by *f*0: a higher tone will make /au/ produced as higher and more front, and vice versa; a rising tone will strengthen the dynamic feature while a falling tone will lead to monophthongization.

Methods. We examined an open-source Standard Mandarin reading text corpus *AISHELL-1*, published by Beijing Shell Shell Technology Co., Ltd. (Bu *et al.* 2017). The data we used were composed of recordings obtained from 10 females and 10 males, which were the same data set used in Li & Al-Tamimi (under review), to make sure that the /ai/ and /au/ data come from the same population and are comparable. The data were automatically segmented and aligned at the syllable (character) and the phoneme levels, using the Montreal Forced Aligner (MFA) (McAuliffe *et al.* 2017). Only lexical tones (1 to 4) in Standard Mandarin are included in the lexicon. The tokens chosen for analysis are from monosyllabic or disyllabic words. Segmental boundaries of the diphthong /au/, the previous and following segments, acoustic information (*f*0 and the first two formants), tonal information, full word information, and utterance information (relative position in the word and in the sentence) were automatically measured and extracted using the automatic script of *Praat* (version 6.3.02) (Boersma 2001). The formant information of all the /au/ items was verified manually. This yielded a data set of 2276 occurrences: 1142 for 10 male speakers and 1134 for 10 female speakers. For each occurrence, we obtained 11 time-normalized intervals, at 10% intervals for formant and *f*0 frequencies.
Since the diphthong /au/ is a dynamic vowel sequence, we used Generalized Additive Mixed Models (GAMMs) (Wood 2006) to capture its dynamic pattern. We performed two sets of modelling. The first model was to evaluate the interaction between the diphthong realization using F1 and F2 dimensions and the tonal unit, which had two categorical predictors: *tone* with four levels, and *sex* with two levels. The second model replaced the tonal information by the combination of *f*0 and *duration*. The outcome in the two models was either F1 or F2 frequencies. We used *time* as a continuous predictor (11 normalized intervals), represented via a *smooth* as a non-linear variable, to track the dynamic pattern during the diphthong realization. The other variables, i.e., the speaker ID, the segmental information, and utterance information, were considered as *factor smooths* modelled as random effects. We then verified the auto-correlation levels of our models and obtained Auto-Regressive GAMMs. The choice of optimal models was done using the functions *gam.check* and *CompareML*.

Results. The results of the modelling generally confirm the hypothesis: the tonal effect on vowel height (related to F1 value) observed on the diphthong /ai/ in Li *et al.* (2023) and Li & Al-Tamimi (under review) also occurs in /au/. More concretely, with a high tone (tone 1), /au/ tends to be realized as more closed; with a low tone (tone 3), it is more open;

with a rising tone (tone 2), it tends to have a typical diphthongized realization, where the F1 contour shows a dynamical pattern; and with a falling tone (tone 4), the diphthong tends to be monophthongized. The predicted F1 contours are shown in Figure 1 (left). As for a more accurate *f*0 – formants interaction, the results show that *f*0 is negatively correlated with F1. For instance, as demonstrated in Figure 1 (right), plotted by *itsadug* package (Van *et al.* 2015), the F1 changes on the z-axis, whereby a higher frequency value of F1 is denoted by the blue end of the color scale, and a lower frequency value of F1 is denoted by the red end of the color scale. The contour lines (which show the same value of F1) and colors show that when *f*0 rises, F1 falls, especially towards the beginning and the ending.
The predictions of F2 by the models are more complicated. The results of the second model show that the assumed positive correlation between *f*0 and F2 only occurs in female cases. As for the tonal effect on vowel backness (related to F2 value), there is no evidence showing the link between tonal height and F2 value. However, the result shows that the degree of dynamic pattern within the diphthong /au/ is related to the tone trend: the F2 contour with a rising tone (tone2) shows a more dynamic pattern than that with a falling tone (tone4 or tone3).

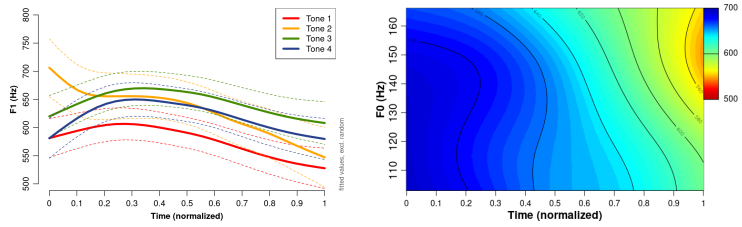


Figure 1: *Results of the modelling. Left: Predicted F1 contours (y-axis) over time (x-axis) of male speakers. Dashed lines represent confidence intervals. Right: Predicted F1 values (different color) with different f0(y-axis) over time (x-axis) of male speakers.*

Discussion. The models established in this study explain the interaction between *f*0/tone and vowel realization in /au/, especially in its second target /u/. The results show that the negative correlation between *f*0 and vowel height (F1) is common in simple vowels of different heights and diphthongs in different directions. This correlation can also be explained by the relationship between *f*0 and larynx height: the vocal folds and vocal tract are related through the physiological linkage between the larynx and hyoid bone/jaw, which is mainly reflected in the vertical direction (Honda *et al.* 1999; Moisik *et al.* 2016; Chen *et al.* 2021). As for the correlation between *f*0 and F2, the results of the model did not fully verify our hypothesis. There are two possible explanations for this. First, the tongue movement direction of /au/ is retracted, and due to physical limitations, the vocal tract space in this direction is limited, which may cause the correlation between *f*0 and vowel backness to be weakened, which is also consistent with (Shaw *et al.* 2016). Secondly, this result may imply that the effect of *f*0 on vowels is mainly reflected in vowel height rather than its backness. That is, the correlation between *f*0 and F2 may be due to the correlation between height and backness in the tongue movement direction of /ai/ in the previous study. More data and analysis are needed to get a more reasonable explanation.

References

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int., 5(9), 341–345*

Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017). AIShell-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline. In *Oriental COCOSDA 2017* (p. Submitted).

Chen, W.-R., Whalen, D. H., & Tiede, M. K. (2021). A dual mechanism for intrinsic f0. *Journal of Phonetics*, 87, 101063.

Erickson, D., Iwata, R., Endo, M., & Fujino, A. (2004). Effect of tone height on jaw and tongue articulation in Mandarin Chinese. In *International symposium on tonal aspects of languages: With emphasis on tone languages.*

Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in f0 control. *Language and Speech*, 42(4), 401–411.

Li, C., Al-Tamimi, J., & Wu, Y. (2023). Tone as a factor influencing the dynamics of diphthong realizations in Standard Mandarin. In *Radek Skarnitzl & Jan Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences (pp. 1876-1880).*

Li, C., Al-Tamimi, J. (under review). The impact of the tonal factor on diphthong realizations in Standard Mandarin: Explanations within the Articulatory Phonology framework modelled via Generalized Additive Mixed Models. *Journal article under review.*

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. interspeech 2017 (pp. 498–502)*.

Moisik, S. R., Lin, H., & Esling, J. H. (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). *Journal of the International Phonetic Association*, 44(1), 21–58.

Shaw, J. A., Chen, W.-r., Proctor, M. I., & Derrick, D. (2016). Influences of tone on vowel articulation in Mandarin Chinese. *Journal of Speech, Language, and Hearing Research*, 59(6), S1566–S1574.

Van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, D. (2015). *itsadug: Interpreting time series and autocorrelated data using GAMMs*.

Wang, P. (2007). *A Statistical Study on the Tones and Vowels of Beijing Dialect*. (Unpublished doctoral dissertation), Nankai University.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic f0 of vowels. *Journal of phonetics*, 23(3), 349–366.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. chapman and hall/CRC

The interaction between phonetics and phonology when processing the acoustic signal: evidence from labial coarticulation in English and French

Phil J. Howson[1], Marianne Pouplier[1], Francesco Rodriquez[1], Eva Reinisch[2], Justin J. H. Lo[3], Chris Carignan[3], Bronwen Evans[3]

[1]Ludwig-Maximilians-Universität München
[2]Austrian Academy of Sciences
[3]University College London
p.howson@phonetik.uni-muenchen.de

Introduction. English and French both have rounded vowels. However, they differ in the fact that English has phonetic lip rounding on back vowels, while French has phonologically contrastive front rounded vowels. So, the difference between these languages in this regard is that French has phonological specification (Dresher, 2009) of a rounding feature for front vowels. The purpose of this study is to examine how the difference in the hypothesized feature specification affects the perception of anticipatory lip rounding.

In English and French anticipatory lip rounding can be both extensive in nature and extremely variable by speaker (Bell-Berti & Harris, 1982; Vaxelaire, Bonnot, & Keller, 1999; Roy, 2005; Noiray et al., 2010; Howson et al., 2021) and both English and French listeners make in principle use of this information to decode the incoming speech signal (Redford et al., 2018; Sock, Hecker, & Cathiard, 1999; Hirsch et al., 2003). This suggests that English and French language users make use of the incoming phonetic information about lip rounding despite differences in the phonological nature of lip rounding in their languages for front vowels. Linguists have long postulated that phonological contrast in one's language enhances perception (e.g., Trubetzkoy, 1939; Boomershine et al., 2008). Therefore, the presence of phonological contrast for roundedness for front vowels in French but not in English opens up the possibility that anticipatory coarticulation for lip rounding is utilized differently for listeners of each language.

Methods. 22 English and 16 French adult participants were recruited for an eyetracking experiment on the basis that their L1 was either American English or Metropolitan French (hereafter, French). Production data were recorded for all participants. To track the lip shape, we used an adapted version of the "blue lip" technique (Lallouache, 1991). Measurements of lip spread were calculated as the distance between the lip corners (Noiray et al., 2011) and temporal differences were quantified using a sigmoid method (Lo et al., 2023). We chose two degrees of anticipatory coarticulation to measure sensitivity to different amounts of anticipatory lip rounding. Tokens were binned into two groups ("extensive" for relatively longer distance coarticulation and "constrained" for relatively shorter distance coarticulation) based on the distribution of the temporal span of coarticulation in the production data. The stimuli for the perception experiment were chosen from a subset of the production data such that the same speakers contributed to both the extensive and constrained coarticulatory conditions. For English and French, 24 & 29 tokens from the "extensive" (1st quartile) and 24 & 29 tokens from the "constrained" (4th quartile) ends of the distribution were chosen for each language respectively. The mean onset of anticipatory lip rounding preceded the vowel target in English by 402 ms for the extensive category and 136 ms for the constrained category. For French, the extensive category mean preceded the target by 288 ms and by 110 ms for constrained.

During the eyetracking experiment, participants saw a minimal pair of words on the screen which had either an unrounded or rounded target vowel (e.g., English: heed / who'd; French: scie / su) and were instructed to click on the target word as soon as they recognized it. Stimuli were presented in the carrier phrase in which they were produced, since anticipatory coarticulation in the production data was quantified over the entire utterance (English: But Tessa had said *target* pleasantly; French: Mais elle déclarait *target* par hazard). English target stimuli had the vowel pairs /i:/ vs. /u:/, /ɪ/ vs. /ʊ/, /eɪ/ vs/ vs. /oʊ/. French target stimuli had /e/ vs. /ø/, /ɛ/ vs. /œ/, /e/ vs. /o/, /ɛ/ vs. /ɔ/, and /i/ vs. /y/.

Incorrect answers were discarded (< 1% of the data) from analysis and looks to the target or competitor were binned at 5ms intervals. Growth Curve Analysis (Mirman et al. 2008) was used to compute the proportion of fixations on the target. The model included a fixed effect for coarticulation (2 levels: extensive, constrained) and functions for time (time^n, n = [1,7]) The interaction between Coarticulation and Time was also included. Random intercepts were included for participant, speaker, and pair (i.e., the pair of words on the screen). We computed one model for English and one for French. To determine differences in growth curves for extensive and constrained conditions, we computed a smoothing spline for each model by randomly sampling participants from each condition and fitting a smoothing curve to their data. This was done 1000 times to generate a distribution and obtain 95% confidence intervals (Wendt et al., 2014).

Results. The results of the GCA for English revealed that there was an increase in target fixations at approximately 75 ms after the onset of the target segment, but no significant difference for the interaction between coarticulation and functions of time (time^n: $p > 0.05$; Figure 1). Given that eye movement planning and execution lags the input stimuli by approximately 200 ms (Travis, 1936), we added 200 ms to the spike in fixations observed in the results to estimate when word recognition took place. English listeners utilized coarticulation at approximately 125 ms before the target vowel onset. The analysis of French, on the other hand, did reveal a significant difference for the interaction between

coarticulation and functions of time (time^n: $p < 0.05$, except time^5: $p = 0.82$) and revealed an increase in fixations towards the target 50 ms before the onset of the critical target in the extensive condition and 110 ms after the onset of the target stimuli in the constrained condition (Figure 1). French listeners had a rapid increase in looks to the target at roughly 200 ms after the onset of coarticulation in both conditions. When taking saccade lag time into account (~200 ms), this suggests listeners recognized the target at approximately 250 ms (extensive condition) and 90 ms (constrained condition) before target vowel onset. This indicates that French listeners use the coarticulation related to lip rounding as soon as it is available in the speech stream, whether that is extensive or more constrained anticipatory coarticulation.

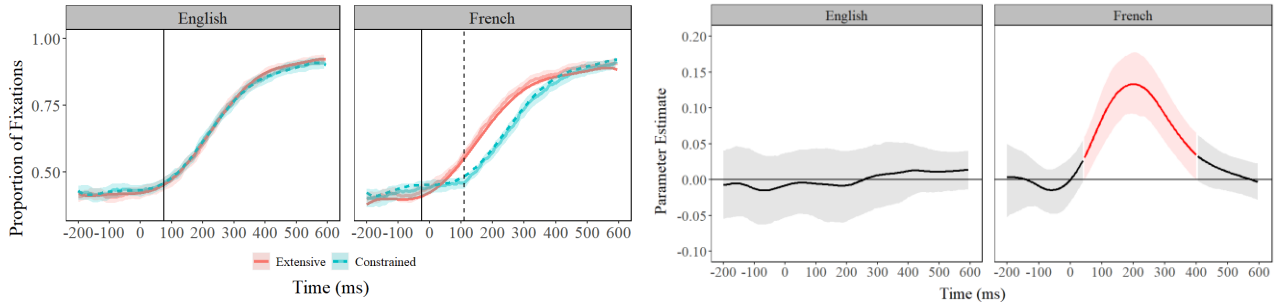


Figure 1 (left): Growth curve analysis for extensive (red) and constrained (blue) for English (left) and French (right). The mean Proportion of Fixations for extensive (opaque red) and constrained (opaque blue) with ±1 SE are also presented. Vertical lines indicate estimated point of increase in fixations on the target for extensive (solid) and constrained (dashed) coarticulation. Figure 1 (right): divergence plots comparing the difference between extensive and constrained for English (left) and French (right). Red indicates, in the right-hand graph, a significant difference between contours. 0 ms indicates the onset of the critical segment ((un)rounded vowel).

Discussion. The results indicated English listeners start to recognize anticipatory lip rounding at around 125 ms before the target vowel onset. The French listeners on the other hand recognize upcoming lip rounding as far back as 250 ms before the target onsets and displayed sensitivity to differences in extensive and constrained coarticulation. The difference in perception contrasts with the differences in the stimuli: English had a mean coarticulatory onset of 402 ms (extensive) and 136 ms (constrained) before the target vowel, while French had a mean onset of 288 ms (extensive) and 110 ms (constrained) before the target vowel. So, despite the availability of coarticulatory information earlier in the speech stream for English than in French, English listeners did not show any differences between their perception of extensive and constrained coarticulation. Additionally, English listeners did not demonstrate an increase in fixations as early as French listeners. The reason for this is possibly due to the phonological status of lip rounding in French. The data thus supports the notion that the presence of a phonological contrast improves sensitivity to subtle differences in coarticulation related to the acoustic-phonetic cues to that contrast. Whether French listeners are equally sensitive to rounding in front-back vowel pairs will have to be addressed in future research.

References

Bell-Berti, F. & Harris, K. S. (1982). Temporal patterns of coarticulation: lip rounding. *Journal of the Acoustical Society of America*, 449-454.

Boomershine, A., Hall, K. C., Hume, E., & Johnson, K. (2008). The impact of allophony versus contrast on speech perception. In Avery P., Dresher E., & Rice K. (eds.), *Contrast in Phonology*, pp. 143-172. Berlin: Mouton de Gruyter.

Dresher, B. E. (2009). The Contrastive Hierarchy in Phonology. Cambridge University Press, Cambridge.

Hirsch, F., Sock, R., Connan, P., & Brock, G. (2003). Auditory effects of anticipatory rounding in relation with vowel height in French. *Proceedings of the International Congress of Phonetic Sciences*, 1445-1448.

Howson, P. J., Kallay, J. E., & Redford, M. A. (2020). A psycholinguistic method for measuring coarticulation in child and adult speech. *Behavior Research Methods*, 846-863.

Lallouache, M. T. (1991). *Un poste visage-parole couleur: Acquisition et traitement automatique des contours des lèvres*. Ph.D. dissertation, Institut National Polytechnique de Grenoble.

Lo, J. J.H., Carignan, C., Pouplier, M., Alderton, R., Rodriquez, F., Evans, B. G., & Reinisch, E. (2023). Language specificity vs speaker variability of anticipatory labial coarticulation in German and English. *Proceedings of the International Congress of Phonetic Sciences*, 2105-2109.

Ménard, L., Cathiard, M., Troille, E., & Girouox, M. (2016). Effects of congenital visual deprivation on the auditory perception of anticipatory labial coarticulation. *Folio Phoniatrica et Logopaedica*, 83-89.

Noiray, A., Cathiard, M., Abry, C., & Ménard, L. (2010). Lip rounding anticipatory control: Crosslinguistically lawful and ontogenetically attuned. In Maassen, B. & van Lieshout, P. (eds.), *Speech motor control: New developments in basic and applied research*, pp. 153-171. Oxford, United Kingdom: Oxford University Press.

Redford, M. A., Kallay, J. E., Bogdanov, S. V., & Vatikiotis-Bateson, E. (2018). Leveraging audiovisual speech perception to measure anticipatory coarticulation. *Journal of the Acoustical Society of America*, 2447-2461.

Roy, J. (2005). Visual perception of anticipatory rounding gestures in French. *Proceedings of Interspeech*, 2949-2952.

Sock, R., Hecker, V., & Cathiard, M. (1999). The perceptual effects of anticipatory labial activity in French. *Proceedings of the International Congress of Phonetic Sciences*, 2057-2060.

Travis, R. C. (1936). The latency and velocity of the eye in saccadic movements. *Psychological Monographs*, 242-249.

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie* [= TCLP 9]. Československo: Praha.

Vaxelaire, B., Sock, R., Bonnot, J., & Keller, D. (1999). Anticipatory labial activity in the production of French rounded vowels. *Proceedings of the International Congress of Phonetic Sciences*, 53-56.

Wendt, D., Brand, T., & Kollmeier, B. (2014). An Eye-Tracking Paradigm for Analyzing the Processing Time of Sentences with Different Linguistic Complexities. *PLoS ONE*, e100186.

Production Allophones of North American English Liquids

Mark Tiede[1], Suzanne Boyce[2], Michael Stern[3], Teja Rebernik[4], Martijn Wieling[4]

[1]Brain Function Lab, Dept. of Psychiatry, Yale University
[2]Dept. of Communication Sciences and Disorders, University of Cincinnati
[3]Dept. of Linguistics, Yale University
[4]Dept. of Information Science, University of Groningen

mark.tiede@yale.edu, boycese@ucmail.uc.edu, michael.stern@yale.edu,
t.rebernik@rug.nl, m.b.wieling@rug.nl

Introduction. The North American English (NAE) syllabic liquids /r/ [ɚ] (as in "purr") and velarized (dark) /l/ [ɘɬ] (as in "pull") form a natural class phonologically and phonetically by traditional acoustic criteria; however, they show a high degree of production variability across speakers (Delattre & Freeman, 1968; Westbury et al., 1998; Mielke et al., 2016). The multiple attested articulatory variants of /r/ in particular converge on a perceptually equivalent acoustic profile with F1 and F2 characteristic of a central vowel and an F3 at 80% or less of the 3rd natural resonating frequency of the vocal tract (Hagiwara, 1995; Espy-Wilson et al., 2000). Laterals are similar but with F3 shifted in the opposite direction. Broadly speaking, both /r/ and /l/ variants have been grouped into tip down ('bunched'/laminal) and tip up ('retroflex'/apical) categories. While some modeling evidence for /r/ suggests F4 differences between these types (Zhou et al., 2008), no perceptual data exist showing that listeners are able to distinguish exemplars of these two production allophones reliably (see e.g. Twist et al. 2007 for a representative null result). Other continuants with production variants typically show consistent acoustics maintained over a smoothly varying range of motor equivalent "trading relations": /u/ for example can be produced with a consistent formant pattern by manipulating the extent of lip protrusion *vs*. laryngeal lowering. /r/ is unusual in that no comparable trading relations exist providing a smooth transition from one postural type to the other, raising questions of how many types exist, how speakers learn their preferred posture, and whether the production goal is driven by an auditory or proprioceptive target. Here we use data scanned using MRI and midsagittal ultrasound from a range of speakers producing NAE syllabic /r/ and /l/, to survey their production variety, and to support a new approach for their categorization.

Methods. 28 native NAE speakers (14F) were scanned with 5 mm slice thickness and 128x128 voxels (1.07 pixel/mm resolution) using midsagittal MRI. Speakers were instructed to produce "purr" or "pull" and to sustain the liquid during the 1.2 s scan duration. Speaker audio recorded immediately prior to and following scanning was used to confirm achievement of the expected acoustic target. Tongue shapes were obtained by fitting a thin plate spline to the surface, from the top of the epiglottis to the anterior-most point of the apex. Distance functions were sampled along a semipolar grid and parameterized as the sum of the first three coefficients from a Fourier transform (Liljencrants, 1971). Unsupervised *k*-means clustering using elbow and silhouette heuristics was used to determine optimal group separation. For additional power an additional 70 Dutch speakers were recorded producing (English) "purr" and "pull" with midsagittal ultrasound using the facilities of SPRAAKLAB (Wieling et al. 2023) during the 2022 Noorderzon Festival (Groningen), of whom 49 were retained following review by native English listeners. Tongue surface contours at the center of the acoustically determined liquid intervals were extracted using DeepEdge (Chen et al., 2020).

Results. Based on *k*-means clustering results, both /r/ and /l/ tongue shapes cluster reliably into three groups driven primarily by tongue dorsum shape. Principal component analysis of tongue shapes showed independently that three components accounted for 99.9% of variance. As shown in Figure 1a, for both liquids these separate into concave, flat and convex patterns, further distinguished by whether the tongue tip is recessed (laminal) or protruded (apical). With the qualification that the tongue tip is not always visible, the same pattern holds for the ultrasound data. This suggests that tongue shapes for NAE syllabic liquids can be characterized using just two parameters: the quadratic term of a polynomial fit to tongue dorsum shape, whose sign and magnitude map onto the degree of convexity/concavity, and a binary feature characterizing recessed (laminal) *vs*. protruded (apical) tongue tip posture. We observed that using the rhotic apical *vs*. laminal tongue tip pattern as a prior predicted the same pattern for the corresponding within-speaker lateral: 83.3% of apical /r/ speakers produced an apical /l/. However, the converse was at chance: 50.0% of apical /l/ speakers produced laminal /r/ (n = 28, MRI only).

Discussion. The extensive variety of observed midsagittal tongue shapes used to produce perceptually equivalent acoustic signatures for /r/ and /l/ likely reflects their interaction with individual differences in speaker palatal morphology. (While misalignment of the sampling plane is also a possibility, shapes were verified against a midsagittal cross-section of coronally-oriented volumes collected during the same session.) Given this variety, how do language learners settle on a preferred shape? Syllabic liquids are notoriously among the last NAE sounds to be acquired, unsurprising given that

they require coordination of at least three constrictions. One possibility may be that children, given sufficient exploration of articulatory possibilities guided by their own perceptual feedback and reinforcement from their parents and peers eventually stumble into a configuration that succeeds in producing the appropriate acoustics. However, a second possibility is that coproduction with other speech targets may expose them to alternative strategies which are close to liquid targets: In two instances participants in this study succeeded in producing separately scanned apical and laminal variants of /r/ with the same acoustics but very different dorsal shapes. Additional scanning of coproduced onset contexts (e.g. "drain" [dr͡eɪn] *vs*. "grain" [gr͡eɪn]) showed an apical posture during the rhotic for the former and a laminal posture for the latter (Figure 1b). Alternative /r/ postures employed by the same speaker have also been found using EMA (Guenther et al., 1999; Tiede et al., 2010) and ultrasound (Mielke et al., 2016). This suggests that fluent NAE speakers have access to more than one production strategy for liquids, selected on least-effort principles during coproduction, but favoring one over others in syllabic contexts as being easier (for them) to produce and sustain. [Work supported by NIH DC05250 and DC002717.]

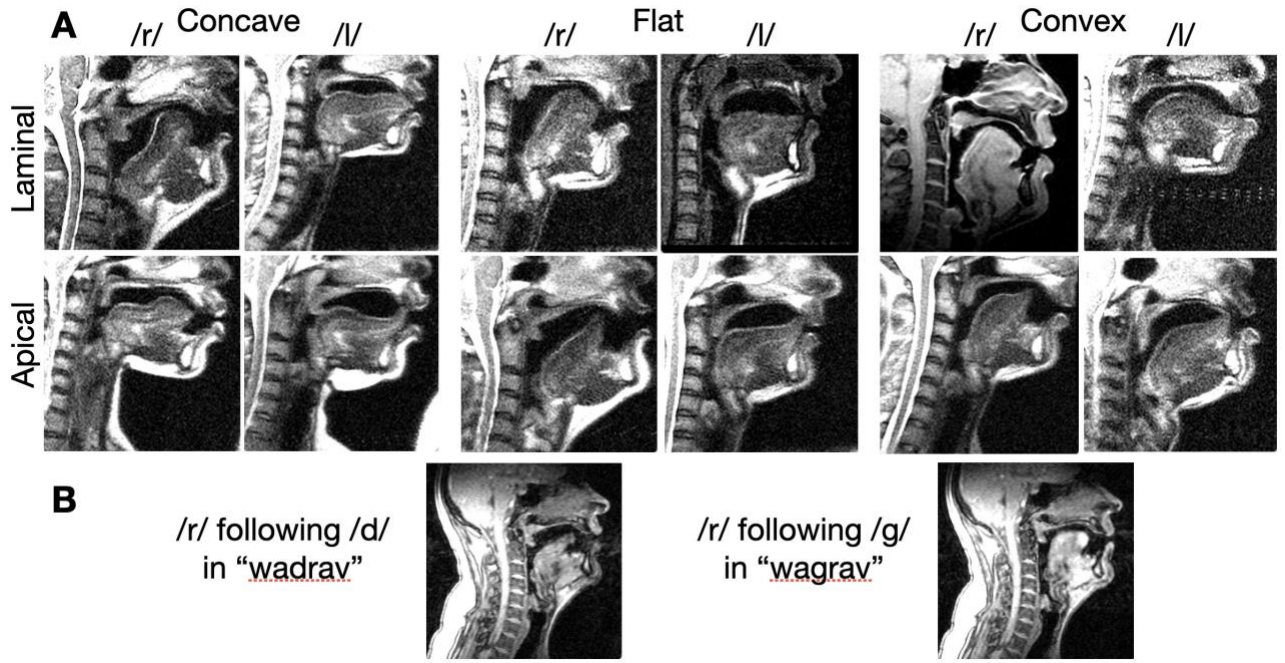


Figure 1: A) *Representative tongue shapes for syllabic liquids showing apical and laminal variants of concave, flat and convex tongue dorsum postures*. B) *Apical (left) vs. Laminal (right) productions of /r/ by the same speaker under differing syllable onset coproduction contexts.*

References

Chen, W-R., Tiede, M., & Whalen, D. (2020). DeepEdge: automatic ultrasound tongue contouring combining a deep neural network and an edge detection algorithm. Paper presented at the *12th International Seminar on Speech Production* (ISSP 2020). https://github.com/WeirongChen/DeepEdge.

Delattre, P. & Freeman, D. (1968) A dialect study of American R's by X-ray motion picture. *Linguistics*, 44, 29-68.

Espy-Wilson, C., Boyce, S., Jackson, M., Narayanan, S. & Alwan, A. (2000) Acoustic modeling of American English /r/. *The Journal of the Acoustical Society of America,* 108(1), 343-356.

Guenther, F., Espy-Wilson, C., Boyce, S., Matthies, M., Zandipour, M., & Perkell, J. (1999) Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *The Journal of the Acoustical Society of America*, 105(5), 2854-2865.

Hagiwara, R. (1995) Acoustic realizations of American English /R/ as produced by women and men. UCLA Working Papers in Phonetics, 90, 1-187

Liljencrants, J. (1971). Fourier series description of the tongue profile. *KTH Speech Transmission Laboratory – Quarterly Progress Status Reports*, 12(4), 9-18.

Mielke, J., Baker, A., & Archangeli, D. (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language*, 101-140.

Tiede, M., Boyce, S., Espy-Wilson, C. & Gracco, V. (2010). Variability of North American English /r/ production in response to palatal perturbation. In Speech Motor Control: New Developments in Basic and Applied Research, 53-67, B. Maassen & P. van Lieshout, Eds. Oxford University Press.

Twist, A., Baker, A., Mielke, J., & Archangeli, D. (2007). Are "covert" /ɹ/ allophones really indistinguishable?. *University of Pennsylvania Working Papers in Linguistics*, 13(2), 207-216.

Westbury, J., Hashi, M., & Lindstrom, M. (1998) Differences among speakers in lingual articulation for American English /ɹ/. *Speech Communication*, 26, 203-226.

Wieling, M., Rebernik, T., & Jacobi, J. (2023). SPRAAKLAB: a mobile laboratory for collecting speech production data. In *Proceedings of the 20th International Congress of Phonetic Sciences* (Prague), 2060-2064.

Zhou, X., Espy-Wilson, C., Boyce, S., Tiede, M. (2008). A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. *The Journal of the Acoustical Society of America*, 123, 4466-4481.

Poster 2

4:30 – 6:30 pm

187	Variability in the articulation of Beijing Mandarin rhotic vowels	Song Jiang (University of Toronto)*; Alexei Kochetov (University of Toronto)
192	Front rounded vowels by English learners of German in read and spontaneous speech	Qiang Xia (Humboldt-Universität zu Berlin); Megumi Terada (Humboldt-Universität zu Berlin)*; Malte Belz (Humboldt-Universität zu Berlin); Christine Mooshammer (Humboldt-Universität zu Berlin)
208	A sound change happening in just one generation: nasal coda lost in the Chengdu variety of Southwestern Mandarin	Sishi Liao (Institute for Phonetics and Speech Processing (IPS), LMU Munich)*; Philip A Hoole (Institute of Phonetics, Munich University); Jonathan Harrington (
101	What does supraglottic articulatory global speed tell us about disfluencies?	Fabrice Hirsch (UMR 5267 Praxiling)*; Ivana Didirková (Université Paris 8 Vincennes - Saint-Denis); Michael Blomgren (University of Utah, Department of Communication Sciences & Disorders); Sofiane Azzouz (UMR 7503 LORIA); Fanny Guitard-Ivent (Praxiling); Slim Ouni (LORIA)
105	Inter-subject variation in tongue shape during vowel production in /b/V/t/ sequence: An rtMRI study using 8 vowels from 74 subjects	Satyadev Badireddi (Indian Institute of Science (IISc), Bangalore)*; Shreya Shrikant Karkun (BMS College of Engineering); Prasanta Kumar Ghosh (Indian Institute of Science (IISc), Bangalore)
99	A pulse-step model of speech motor control: Evidence for an extrinsic pacemaker	Alan Wrench (Articulate Instruments/Queen Margaret University)*
87	Auditory Feedback Perturbation of F2 in French-speaking Children	Isabelle Démosthènes (Université du Québec à Montréal)*; Lucie Ménard (Université du Québec à Montréal)
96	The role of executive functions and levodopa on articulatory timing	Elisa Herbig (University of Cologne)*; Tabea Thies (University of Cologne); Michael Barbe (University Hospital Cologne); Doris Muecke (ifL Phonetics - University of Cologne)
163	Effects of Pharyngealization and Labialization on Formants in Tashlhiyt	Philipp Buech (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne Nouvelle)*; Anne Hermes (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS & Sorbonne Nouvelle, Paris); Rachid Ridouane (LPP (CNRS & Sorbonne Nouvelle))
139	On the Supra-laryngeal Articulation of Prosodic Prominence in Southwestern Mandarin	Jing Huang (National Tsing Hua University)*; Feng-fan Hsieh (National Tsing Hua University); yueh-chin chang (NTHU)
170	Human Tongue Finite Element Model Validation with 3D MRI of subject specific vowel articulations	Maxime Calka (ISCD, Sorbonne Université)*; Pascal Perrier (Gipsa-lab, Grenoble INP, Université Grenoble Alpes); Yohan PAYAN (Univ. Grenoble Alpes)
70	Variability in Czech children's sibilant fricative production	Tanja Kocjančič Antolík (Charles University)*; Kateřina Vitásková (Palacký University Olomouc); Kateřina Bujoková (Charles University); Tomáš Bořil (Charles University)
45	The role of the supplementary motor area in speech production: Evidence from participants who do, and do not stutter.	Charlotte E. E. Wiltshire (Bangor University)*; Nicole Benker (Institute of Phonetics, Munich University); Anton Gadringer (Institute of Phonetics, Munich University); Rosa Hufschmidt (Institute of Phonetics, Munich University); Philip A Hoole (Institute of Phonetics, Munich University)
11	Task effects and phonological error patterns in Australian English-Dutch bilingual children	Hayo Terband (Department of Communication Sciences and Disorders, University of Iowa)*; Bhavana Bhat (Department of Communication Sciences and Disorders, University of Iowa); Anniek Van Doornik (HU University of Applied Science)
116	Compensatory Strategies in Individuals with Moebius Syndrome: A Case Study	Anne Hermes (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS & Sorbonne Nouvelle, Paris)*; Ivana Didirková (Université Paris 8 Vincennes - Saint-Denis); Philipp Buech (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne Nouvelle); Gilles Vannuscorps (Université catholique de Louvain)
233	Validating the Use of Simulation Based Inference for Feedback Aware Control of Tasks in Speech (FACTS) and Human F1 Compensation Data	Alvince L Pongos (Berkeley)*; Kwang S Kim (Purdue University); Ben Parrell (University of Wisconsin-Madison); Vikram Ramanarayanan (University of California, San Francisco & Modality.AI); Jessica L Gaines (UC Berkeley - UCSF Graduate Program in Bioengineering); Srikantan Nagarajan (UCSF); John F Houde (University of California San Francisco)
135	Model simulations suggest that speech motor control is more sensitive to estimated than true sensory noise levels	Jessica L Gaines (UC Berkeley - UCSF Graduate Program in Bioengineering)*; Kwang S Kim (Purdue University); Ben Parrell (University of Wisconsin-Madison); Vikram Ramanarayanan (University of California, San Francisco & Modality.AI); Alvince L Pongos (Berkeley); Srikantan Nagarajan (UCSF); John F Houde (University of California San Francisco)
52	An automated pipeline for preprocessing spontaneous L2 English prosody	Sylvain Coulange (Université Grenoble Alpes)*; Tsuneo Kato (Doshisha University); Solange Rossato (Univ. Grenoble Alpes); Monica Masperi (Univ. Grenoble Alpes)
132	Discovering phoneme-specific critical articulators through a data driven approach	Jesuraj Bandekar (IISc)*; Sathvik Udupa (Indian Institute of Science); Prasanta Kumar Ghosh (Indian Institute of Science (IISc), Bangalore)
201	Cerebellar degeneration impairs adaptation to pitch perturbations in sustained vocalization	Anneke W. Slis (University of Wisconsin-Madison)*; Ben Parrell (University of Wisconsin-Madison)
80	Speech sensorimotor adaptation in individuals with hearing-impairment	Monica Ashokumar (Univ. Grenoble Alpes, GIPSA - Lab)*; Jean-Luc Schwartz (GIPSA-lab); Takayuki Ito (GIPSA-lab)
103 (Remote)	Influence of stress and sequence position on vowel sandhi in Brazilian Portuguese	João Paulo Moraes Lima dos Santos (Universidad de Salamanca)*
181	Phonetic accuracy in French learners of English: towards a bilingual database combining articulatory MRI and audio	Alice Léger (Université Paris Cité)*; Coline Caillol (Université Paris Cité); Emmanuel Ferragne (Université Paris Cité); Hannah King (Université Paris Cité); Sylvain Charron (Université Paris Cité); Clément Debacker (Université Paris Cité); Maliesse Lui (Université Paris Cité); Catherine Oppenheim (Université Paris Cité)

109	Spatio-temporal properties of Japanese coronal consonants: An ultrasound study of /d/ and /r/	Maho Morimoto (Sophia University/JSPS)*; Takayuki Nagamine (Lancaster University)
220	Comparing the real-time perception of French nasal and labial coarticulation	Francesco Rodriquez (Ludwig-Maximilians-Universität München)*; Marianne Pouplier (LMU); Phil J Howson (Ludwig-Maximilians-Universität München); Eva Reinisch (Austrian Academy of Sciences); Justin J.H. Lo (Lancaster University); Christopher Carignan (University College London); Bronwen G. Evans (University College London)
240	Relationship between working memory and auditory rhythm discrimination in adults who stutter	Emily Garnett (University of Michigan)*; Bailey Rann (Michigan State University); Nicholas Mularoni (University of Michigan); Eora Smith (Michigan State University); Soo-Eun Chang (University of Michigan); Devin McAuley (Michigan State University)
54	Identifying different types of lingual tremor in individuals with Parkinson's disease using electromagnetic articulography: a follow-up study	Teja Rebernik (University of Groningen)*; Hidde Jacobi (University of Groningen); Mark Tiede (Yale University); Martijn Wieling (University of Groningen)
120	Speech Rhythm as a Coordinative System Stabilizing Speech Production in Auditory Feedback Perturbations	Bingyu Li (Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle))*
58	Articulatory timing in Hindi CV sequences	Shihao Du (Universität Potsdam)*; Indranil Dutta (Jadavpur University); Adamantios Gafos (U of Potsdam, Department Linguistik)
207	What's in a name? Production (and Perception) of Difficult to Pronounce Names in Academic Settings	Michelle Middaugh Cifuentes (University of Alberta); Daniel Pape (McMaster University)*
38	Association between Speech Motor Learning and Model-based Estimates of Memory Retrieval	Thomas Wilschut (University of Groningen); Katharina M. Pohlener (University of Groningen); Thomas B Tienkamp (University of Groningen); Hedderik van Rijn (University of Groningen); Catharina Sibert (University of Groningen); Defne Abur (University of Groningen)*
162 (Remote)	Articulatory and vocal speaker variability in connected speech	Maria Mendes Cardoso (Federal University of Minas Gerais)*; Adelino Pinheiro Silva (Police Academy of Minas Gerais)
184	Beyond speech production: sensorimotor contribution to native and non-native phoneme perception	Tsuyi Tseng (LDL, Université Lumière Lyon 2)*; Jennifer Krzonowski (LDL, Université Lumière Lyon 2); Alice C. Roy (LDL, Université Lumière Lyon 2); Claudio Brozzoli (ImpAct team, Centre de Recherche en Neurosciences de Lyon); Véronique Boulenger (LDL, CNRS)
209	A dynamic nasalance analysis of /hi/ in Zuberoan Basque	Ander Egurtzegi (CNRS-IKER)*; Andrea Gard-a-Bravo (IPS-LMU Munich, IKER-UMR5478, UPPA); Iñigo Urrestarazu-Porta (CNRS-IKER, UPPA, UPV/EHU)
193	Effects of phonetic contexts on aerodynamic conditions for uvular trills in French	Andres F Lara (LPP)*; Didier Demolin (LPP CNRS); Claire Pillot-Loiseau (Sorbonne Nouvelle University)
24	Relating frication to articulation in Standard Mandarin apical vowels	Sean Foley (University of Southern California)*; Bowei Shao (Département d'études cognitives, Ecole Normale Supérieure - Université PSL); Matthew Faytak (University at Buffalo)
165	Parametric Excitation of Vocal Tract Resonances by Vocal Fold Motion: A Source-Excitation-Filter Model of Speech Production	Gordon Ramsay (Emory University)*
5	Effect of following vowel context on Sevillian derived stop-h sequences	Madeline Gilbert (Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle))*
65	Sex-specific patterns in intraoral pressure in the production of Georgian and German ejectives	Nato Sulaberidze (Friedrich Schiller University Jena)*; Adrian P. Simpson (Friedrich Schiller University Jena)
143	Discovering dynamical models of articulatory gestures from data	Sam Kirkham (Lancaster University)*
228	Spatiotemporal Coupling of the Jaw and Lower Lip: Comparing Talkers with Parkinson's Disease and Amyotrophic Lateral Sclerosis	Mili Kuruvilla-Dugdale (University of Iowa)*; Antje Mefferd (Vanderbilt University Medical Center)
130	Effect of neurotype identity of conversational partners on speech behavior and communicative success	James G Taylor (University of Oregon)*; Melissa A Redford (University of Oregon)
107	Palatalization in Russian fricatives	Natalja Ulrich (University of Oulu)*; Jalal Al-Tamimi (Université Paris Cité)
21 (Remote)	Perception of Accentual Phrase Boundaries in L1 and L2 French	Caroline L Smith (University of New Mexico)*; Bruno Pinto Silva (University of New Mexico)
250	Sexual dimorphism of vocal tract development from pregnancy to adulthood: Mixed-effects modelling of an extended X-ray database	Guillaume Barbier (GIPSA-Lab, Univ. Grenoble Alpes); Louis-Jean Boë (GIPSA-Lab, Univ. Grenoble Alpes); Guillaume Captier (Anatomy Laboratory, Univ. Montpellier); Rafael Laboissière (CNRS Grenoble Alpes University, France)*
199	Combining manual control of intonation with whisper articulation in voice substitution: the case of contrastive focus	Delphine Charuau (GIPSA Lab)*; Nathalie Henrich Bernardoni (CNRS); Silvain Gerber (CNRS); Olivier Perrotin (CNRS)
150	Laryngeal movements in the production of French stops, with variations in phonation mode and intensity	Maeva GARNIER (GIPSA-lab)*; Myriam Fanzone (CHU Nîmes Caremeau); Nathalie Henrich Bernardoni (CNRS)

Variability in the articulation of Beijing Mandarin rhotic vowels

Song Jiang[1], *Alexei Kochetov*[1]

[1]University of Toronto

soong.jiang@mail.utoronto.ca, al.kochetov@utoronto.ca

Introduction. One of the most documented characteristics of the North American English rhotic /ɹ/ is its contextual and/or inter-speaker variability in the choice of tongue shapes – bunched or retroflex (Delattre & Freeman 1968, Mielke *et al.* 2016, among others). In contrast, the situation with Mandarin rhotic vowels (underlying or *er*-suffixed) is much less clear. Some articulatory (EMA or ultrasound) studies of Beijing Mandarin (BM) rhotic vowels (e.g., Lee 2005) reported exclusively tip-down (bunched) articulations; others found either consistently or predominantly tip-up (retroflex) tongue shapes (Xing 2022). Crucially, none of the studies have reported vowel-specific variability on in tongue shapes, similar to that reported for the English rhotic.

To further explore the individual and contextual variability, we are conducting a systematic ultrasound investigation of various vowel qualities in BM – rhotic and non-rhotic. As the data collection is now ongoing, here we are presenting preliminary results based on six speakers.

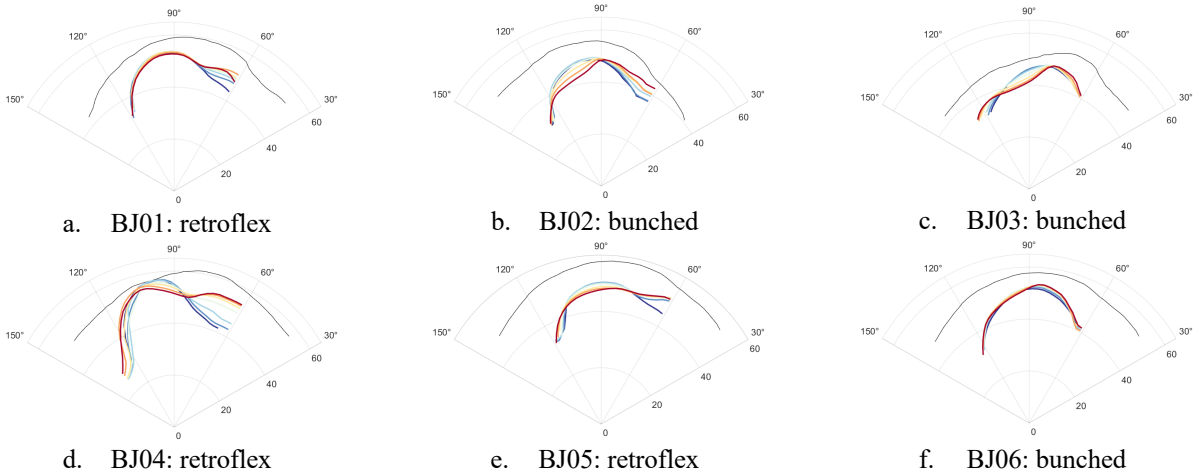
Methods. Ultrasound data were collected from six BM speakers (4 females) at the University of Toronto phonetics lab using an EchoB system (Articulate Instruments Ltd.), set at a frame rate of 60 fps. An UltraFit headset was used to stabilize the probe during imaging. Audio-ultrasound synchronization was implemented in AAA software.

The stimuli were comprised of meaningful words with the vowels /u, a, ə/ and their *er*-suffixed counterparts [u˞, aɚ, ɚ] preceded by bilabial stops (e.g. [pu] 'no' - [pu˞] 'step-dim'; [pa] 'tyranize' - [paɚ] 'handle-dim'; [pʰən] 'gush' - [pʰɚ] 'basin-dim'). The participants produced the target words in the carrier phrase "__, *mà* __ *ba*" ("__, curse with the word __") five times.

Tongue contours were traced using the DeepLabCut method within AAA. For each acoustically defined rhyme, seven equally timed frames were extracted (further referred to as *t1-t7*) and converted to polar coordinates. Polar Smoothing Spline ANOVAs (SSANOVAs) were used to compare the tongue shapes at different points within a rhyme or between different rhymes.

Results.

(i) Individual variation: Figure 1 shows the dynamic change of tongue shape of [u˞]. The results revealed individual variation in the articulation of rhotic vowels. Three participants BJ02, BJ03, and BJ06 used a 'tip-down' tongue shape to produce [u˞]: the tongue blade was raised, and a concavity created in the dorsal region. The other three BJ01, BJ04, and BJ05 used a retroflex configuration; the tongue tip was raised, and the tongue dorsum maintained an [u]-shape.

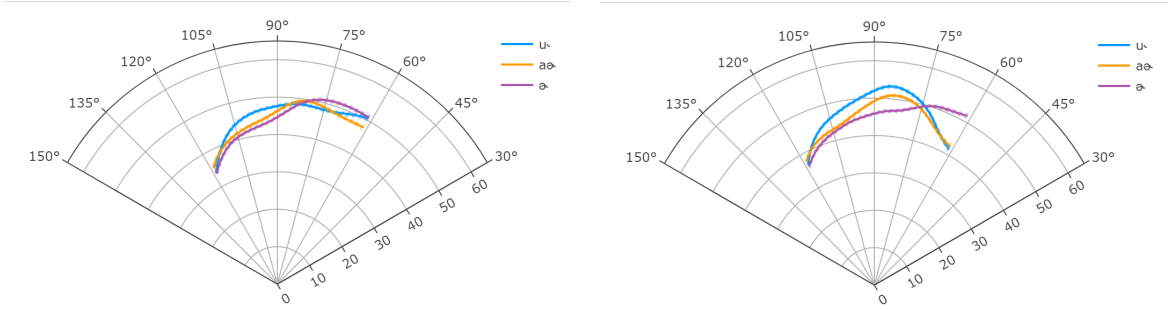


a. BJ01: retroflex b. BJ02: bunched c. BJ03: bunched

d. BJ04: retroflex e. BJ05: retroflex f. BJ06: bunched

Figure 1: *Dynamic changes for the tongue shape of [u˞] (Blue line: t1-the beginning; red line: t7-the end)*

(ii) Contextual variation: Figure 2 shows a within-speaker comparison of the tongue contours at the end of the rhyme in various contexts**.** As can be seen in (a), BJ05 used a retroflex configuration for the articulation of [uɻ], whereas using a bunched tongue shape for [aɚ] and [ɚ], with the tongue body having a concave shape compared to the [uɻ]'s convex dorsum. BJ06 in (b), on the other hand, used a retroflex tongue shape for [ɚ] but not for [uɻ] and [aɚ], with the tongue body being bunched up and the tongue tip pointing down. The two speakers' vowel-specific strategies are therefore not the same. Interestingly, the other four speakers showed consistency in lingual configurations regardless of the vowel.



a. BJ05: only retroflex for [uɻ] b. BJ06: only retroflex for [ɚ]

Figure 2: *SSANOVAs of the t7 tongue contours of [uɻ, aɚ, ɚ] (dashed lines: 95% confidence intervals)*

(iii) Tongue position overlapping: The tongue contours were much less spaced out for the rhotic vowels compared to their non-rhotic counterparts. Figure 3 shows the SSANOVA plots for the mid-point (*t4*) of the non-rhotic vowels and the end point (*t7*) of the *er*-suffixed ones from one retroflexing speaker and one bunching speaker. For both speakers, either the tongue tip or the tongue blade was raised, while the tongue body was lowered, resulting in similar tongue positions.

a. BJ01: retroflex b. BJ02: bunched

Figure 3: *SSANOVAs of the non-rhotic vowels [u, a, ə] (left) and the corresponding er-suffixed form (right)*

Discussion. Preliminary results from six speakers show both individual and contextual variation in the production of BM rhotic vowels. First, our speakers varied in using either a retroflex or a bunched configuration. Neither of these configurations were dominantly used by our participants. Second, we also found some within-speaker variation conditioned by vocalic contexts, albeit not systematically. This, nevertheless, is notable, as previous studies of BM rhotics assumed a contextual uniformity of tongue shapes, highlighting the difference in this respect from the English rhotic. Third, our data also showed that rhotic vowels in BM tend to be more similar to each other compared to their non-rhotic counterparts, which is consistent with the finding in other languages such as Kalasha (Hussain & Mielke 2021). Overall, these results demonstrate considerable variability in the production of BM rhotic vowels even within a relatively small sample of speakers. This highlights the need for a larger-scale investigation of rhotic sounds in Beijing Mandarin (as well as across languages), with these sounds produced in a variety of phonetic contexts and lexical items.

References

Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics, 6*(44), 29–68.

Hussain, Q., & Mielke, J. (2021). An acoustic and articulatory study of rhotic and rhotic-nasal vowels of Kalasha. *Journal of Phonetics, 87*, 101028.

Lee, W.-S. (2005). A phonetic study of the "er-hua" rimes in Beijing Mandarin. *In Ninth European Conference on Speech Communication and Technology*, 1093–1096.

Mielke, J., Baker, A., & Archangeli, D. (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language, 92*(1), 101-140.

Xing, K. (2021) *Phonetic and phonological perspectives on rhoticity in Mandarin.* [Doctoral dissertation: The University of Manchester]

Front rounded vowels by English learners of German in read and spontaneous speech

Qiang Xia, Megumi Terada, Malte Belz, Christine Mooshammer

Department of German Studies and Linguistics, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

qiang.xia.1@hu-berlin.de, megumi.terada@hu-berlin.de

Introduction. The acquisition of front rounded vowels /yː øː/ in German presents a challenge for second language (L2) speakers whose native language (L1) lacks these phonemes. British English, for example, has only two rounded long vowels, /uː ɔː/ (Roach 2004), while German has a different set of rounded long vowels, /yː øː uː oː/ (Kohler 1990). Recently, studies have shown that orthographic representation can influence speech perception and production (cf. Hayes-Harb and Barrios 2021). In German, /yː øː/ are represented orthographically by <ü ö>, respectively. Thus, the question arises as to whether orthography assists L2 learners in producing front rounded vowels in a more target-like manner than in spontaneous speech, where this information is not readily available. Regarding the phonological description, the rounded vowel pairs /yː – uː/ and /øː – oː/ should have the same degree of tongue height (F1) and lip rounding (F3), but differ in tongue advancement (F2). Raphael et al. (1979) indicated that lip-rounding would lengthen the vocal tract, leading to a lowering effect on F2 and F3. Here, we compare the effect of register (read vs. spontaneous speech) on the realisation of front rounded vowels in L2 German. We hypothesise that if orthographic forms are available, the realisation of /yː øː/ will be more front (higher F2) and more rounded (lower F3) than in spontaneous speech because the distinction is marked in orthography.

Methods. We use a subcorpus from the Corpus of Non-Native Addressee Register (Lüdeling et al. 2023), consisting of four English learners of German with self-reported proficiency levels between B1 and C1 according to the Common European Framework of Reference for Languages (CEFR). Each L2 speaker repeated the experiment five times with different L1 interlocutors. The procedure involved a read word list at the beginning, two spontaneous task-based conversations (Baker and Hazan 2011, Diapix task), a task-free spontaneous conversation, and a second reading of the word list at the end of the experiment. The word lists contained all 15 German vowels described in Kohler (1990), embedded in carrier sentences ('Say X please'). Each L2 speaker read the word lists without their L1 conversational partner present. In the Diapix conversation, speakers had eight minutes to find differences between two pictures. All read vowels in the word list were labelled in Praat. In the Diapix data, the vowels /iː eː yː øː uː oː/ were labelled in stressed and accented position in content words (/iː eː/ as a control group for close front unrounded vowels). The data was converted into an EMU database (Winkelmann et al. 2020) and corrected for formant trajectories before further analysis in R (R Core Team 2023). Two native German annotators listened to all annotated vowels and labelled them with "1" or "0" for target-like and non-target-like realisations. Formants were obtained by extracting and averaging five samples around the midpoint of the vowel. The mean formants were normalised to make them comparable across speakers (Lobanov 1971). To assess the learners' proficiency in German, their accent was rated in a follow-up perception study by 27 German native raters on a seven-point Likert scale ranging from no accent (1) to strong accent (7).

Results. In read speech, 80.9% of the 434 vowels were articulated in a target-like way, while the percentage in spontaneous speech is slightly lower, 78.1% of 975. The speaker with the highest accent rating of 5.44 realised only 56.4% of all vowels target-like in spontaneous speech, but reached a higher accuracy of 72.2% in word list. In total, 297 of the total 1409 vowels were assessed as non-target-like and hence excluded in the analysis. We find that front rounded vowels in read speech (lst) show a higher F2 than in spontaneous speech (pix), see Figure 1. F3 does not show any systematic visual effects. However, the orthographic representation seems to help the less proficient speaker 5.44 in rounding /yː/ (M_lst = -1.20, M_pix = -0.06, F(69) = -5.63, p < .001), while her realisation in spontaneous speech resembles [iː] (M_i = -0.74, F(117) = 0.67, p = 0.13), i.e, no significant differences in rounding feature.

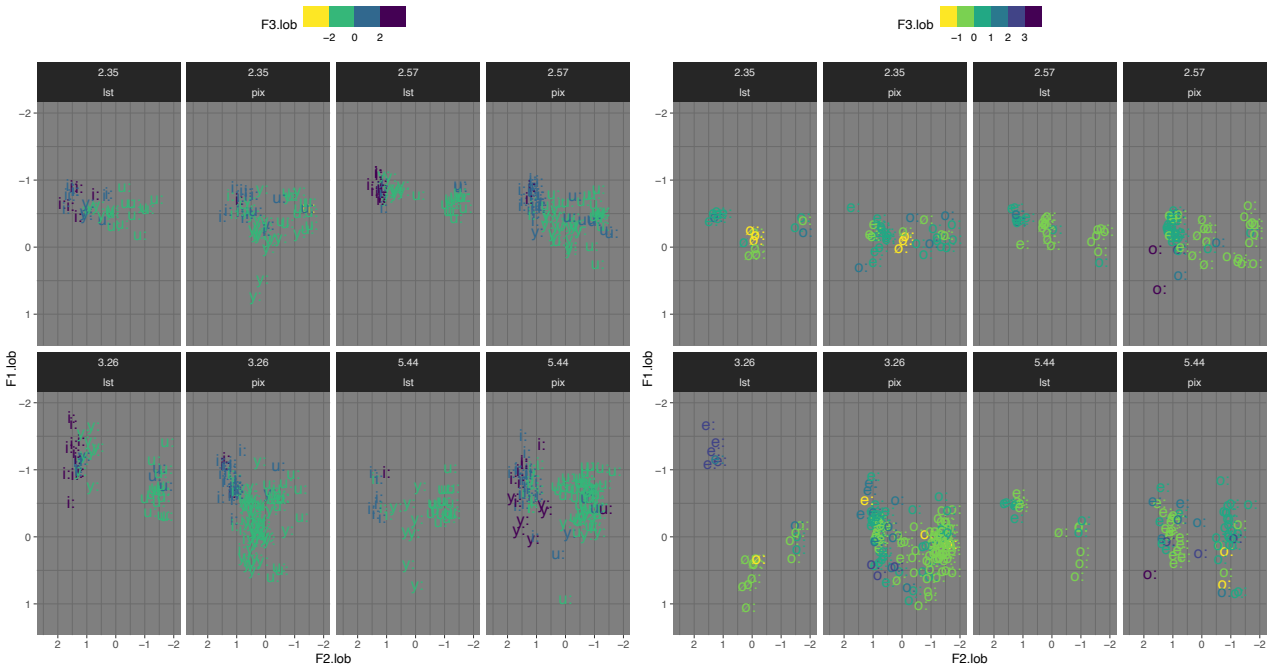


Figure 1: Close (left) and mid-close vowels (right) by English learners of German in read speech (lst) and spontaneous speech (pix). Speakers are indexed with their mean accent ratings. F3 is shown in color (the brighter, the more rounded.)

Discussion. The realisation of front rounded vowels by L2 German speakers is affected by both the availability of orthographic cues in the read register and the learners' proficiency level. The more distinct and dispersed nature of vowels in read speech could be explained by hyperarticulation (Lindblom 1990), potentially cued by the sentence stress of the target word in the list. Although we aimed for vowel tokens in pitch-accented syllables in spontaneous speech, it is possible that prosodically the degree of accent was more variable and sometimes smaller than in read speech due to longer utterances and different positions in the utterance. The larger variability and greater overlap between vowel categories in spontaneous speech can also be explained by more extensive coarticulation due to the varying contexts.

Whether the register effects observed in L2 speakers are comparable to that in L1 speakers will be investigated in the near future. By comparing the realisations of L1 and L2 speakers we will be able to distinguish between the role of the visual availability of the orthographic representation and register effects for learners and native speakers.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334. We thank our student assistant Torben Schilling for his annotation work.

References.

Baker, Rachel and Valerie Hazan (2011). "DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs". In: *Behavior Research Methods* 43 (3), pp. 761–770. DOI: 10.3758/s13428-011-0075-y.

Hayes-Harb, Rachel and Shannon Barrios (2021). "The influence of orthography in second language phonological acquisition". In: *Language Teaching* 54.3, pp. 297–326. DOI: 10.1017/S0261444820000658.

Kohler, Klaus (1990). "German". In: *Journal of the International Phonetic Association* 20.1, pp. 48–50. DOI: 10.1017/S0025100300004084.

Lindblom, Björn (1990). "Explaining Phonetic Variation: A Sketch of the H&H Theory". In: *Speech Production and Speech Modelling*. Ed. by William J. Hardcastle and Alain Marchal. NATO ASI Series. Dordrecht: Springer, pp. 403–439. DOI: 10.1007/978-94-009-2037-8_16.

Lobanov, Boris M. (1971). "Classification of Russian vowels spoken by different speakers". In: *JASA* 49.2B, pp. 606–608.

Lüdeling, Anke, Christine Mooshammer, Robert Lange, Bianca Sell, and Megumi Terada (2023). "Corpus of Non-Native Addressee Register (CoNNAR): Version 1". In: *Media repository of Humboldt-Universität zu Berlin*. URL: https://rs.cms.hu-berlin.de/phon.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Raphael, Lawrence J., Fredericka Bell-Berti, René Collier, and Thomas Baer (1979). "Tongue Position in Rounded and Unrounded Front Vowel Pairs". In: *Language and Speech* 22.1, pp. 37–48. DOI: 10.1177/002383097902200103.

Roach, Peter (2004). "British English: Received Pronunciation". In: *JIPA* 34.2, pp. 239–245. DOI: 10.1017/S0025100304001768.

Winkelmann, Raphael, Klaus Jaensch, Steve Cassidy, and Jonathan Harrington (2020). *emuR: Main Package of the EMU Speech Database Management System*.

A sound change happening in just one generation: nasal coda lost in the Chengdu variety of Southwestern Mandarin

Sishi Liao, Phil Hoole, Jonathan Harrington

Institute for Phonetics & Speech Processing, LMU Munich
sishi.liao|hoole|jmh@phonetik.uni-muenchen.de

Introduction. Anticipatory nasalization in the vowel-nasal (VN) sequence often leads to the reduction of the nasal coda consonant (Ohala & Busa, 1995), resulting in either a nasal vowel Ṽ (Rochet, 1976) or a shifted oral vowel V (Cresci, 2019). It has been reported that the nasal coda in /(V)an/-rime words is lost in the Chengdu variety of Southwestern Mandarin (Liao et al., 2022), combined with a raising and fronting in the pre-nasal vowel quality (Liao et al., 2023). In this study, we try to investigate this sound change with data from two generations (speakers in their 20s and 50s), and we try to figure out whether this sound change is complete or ongoing, and whether any gender difference shows up.

Methods. This study recruited 27 native Chengdu speakers from sex balanced old/young age groups. The mean age for the older group was 58.93 years, and for the younger group 23.08 years. Each participant was recorded with a nasalance device, separating speech signals from the oral and the nasal cavity. The speech materials consist of single words in C(G)V(N)-T structure, with a glide (G) in /∅, j, w/, a rime in V(N) structure /a, an, aŋ/, and an initial consonant (C) in /t, t^h, p, p^h/ to ensure an equal distribution across the 4 tonal categories.
The data from 20 speakers (a total of 20 speakers * 3 rimes * 3 glide types * 4 tones = 720 tokens) were analyzed. The amplitude of the nasal and oral channels (A_n and A_o) was extracted for the sonorant interval (the final (G)VN)). The nasalance score was calculated from $A_n / (A_n + A_o)$, which involved some modification on the Horii Oral-Nasal Coupling Index (Horii, 1980). For each rime, the nasalance score was resampled to 100 datapoints, lowess smoothed with a fraction of 0.3 and was then put into the discrete cosine transform (DCT) in order to measure the shape of the nasalance trajectory. The resulting DCT coefficients k_0 and k_2 are proportional to the mean and curvature respectively (Harrington et al., 2008; Watson & Harrington, 1999). The orthogonal projection (*op*) ratio was calculated by speaker, in order to determine the relative position of tokens to the oral rime /a/ and to the nasal rime /aŋ/: values closer to -1 indicate a token closer to oral rime /a/, and +1 closer to nasal rime /aŋ/.
Linear mixed-effect models were applied to the *op* ratios to test the proximity towards the /a, aŋ/-tokens. The *op* ratio of each observation was set as the response, the `RIME/SPEAKER_GROUP` as the fixed factor, and the `SPEAKER`, `GLIDE` and `TONE` as the random factors. The model was then applied with ANOVA to test the difference between /a, an/-rimes within `SPEAKER_GROUP` and the difference of /an/-rime among speaker groups (and sex, age).

Results. The nasalance score as a function of time for each speaker group is shown in the top-left panel in Figure 1, with the respective DCT coefficients $k_0 \times k_2$ space attached on the top-right panel. The top two panels show that the nasalance score of /an/-rime for older speakers falls between the oral /a/-rime and the nasal /aŋ/-rime, while for the younger speakers the orality of /an/-rime extends towards the oral /a/-rime, with some even reaching beyond.
The orthogonal projection ratio for the /a, an, aŋ/-rimes for each speaker group is shown in the bottom panel of Figure 1, from which we can observe an increase of proximity between /a, an/-rimes from left (old male speakers) to right (young female speakers).
The statistical analysis with *op* ratio as the dependent variable showed a significant difference between /a, an/-rimes for old male ($p < 0.05$), old female ($p < 0.001$), a not quite significant difference in young male ($p = 0.07$), and no significant difference for young female speakers ($p = 0.15$). The difference is significant in /an/-rime nasalance among groups of speakers ($p < 0.001$): in both sex ($p < 0.05$) and age group ($p < 0.001$).

Discussion. The results of this study are consistent with a sound change in progress by which there is an ongoing reduction of nasalization /an/-rime in the Chengdu variety of Southwestern Mandarin. The nasal consonant in /an/-rime is lost and becoming increasingly oral with the pre-nasal vowel being raised and fronted (Liao et al., 2023). The findings confirm the phonologization of /an/-rime de-nasalization (Liao et al., 2022), and are compatible research on other languages and dialect varieties (Busà, 2003; Hajek & Maeda, 2000; Ohala & Busa, 1995) showing a phonetically motivated nasal loss.

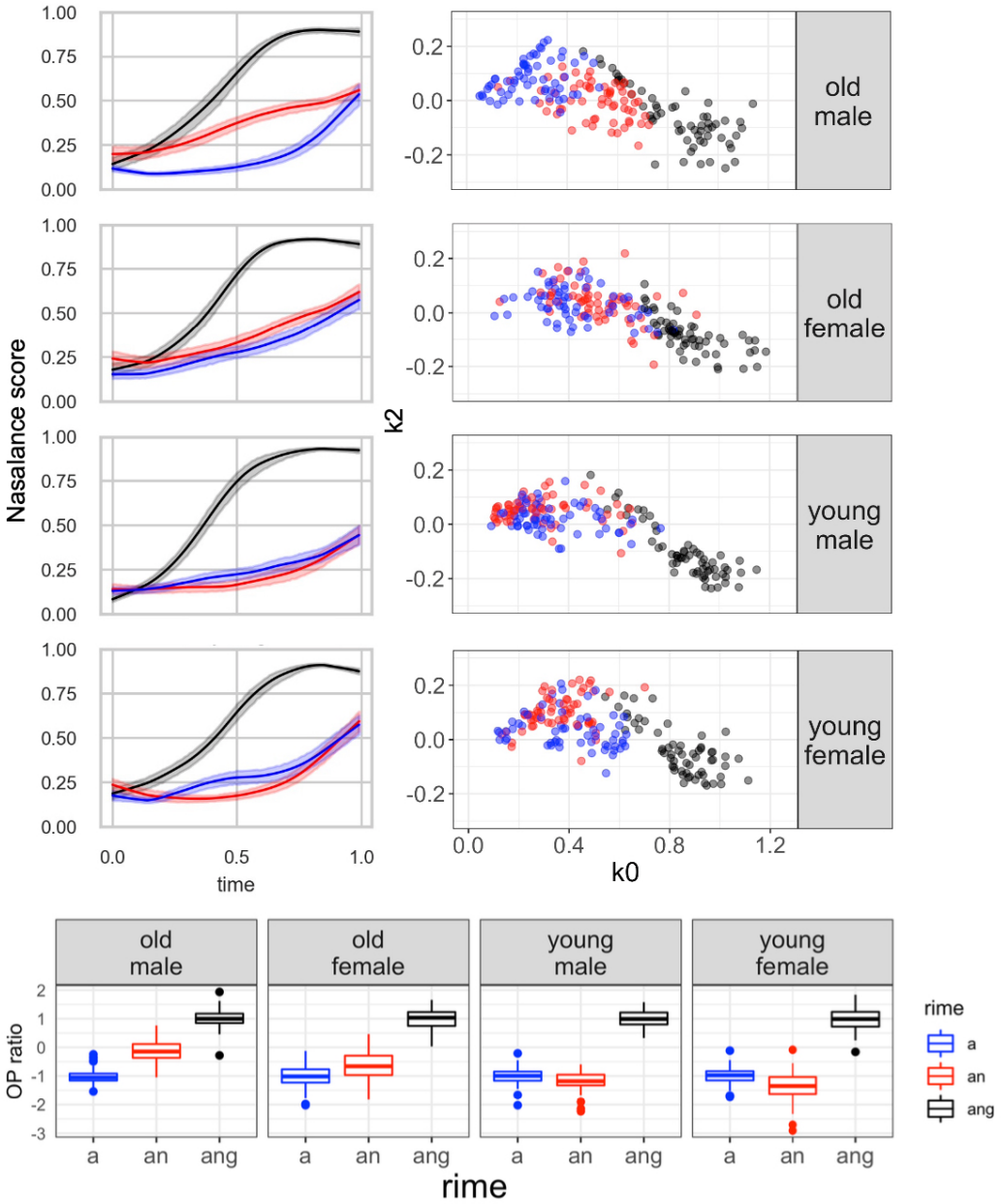


Figure 1. *The nasalance score as a function of time (top-left) and the DCT coefficients $k_0 \times k_2$ space (top-right) for each speaker group. The orthogonal projection ratio to the line connecting the centroids of speaker-specific /a, aŋ/-rimes (bottom panel). The legend colors apply to all plots.*

References

Busà, M. G. (2003). Vowel Nasalization and Nasal Loss in Italian. *ICPhS*, 711–714.

Cresci, M. (2019). VN > V in Camuno: An alternative historical pathway to nasal loss. *Italian Journal of Linguistics*, *31*(1), 61–92. https://doi.org/10.26346/1120-2726-132

Hajek, J., & Maeda, S. (2000). Investigating universals of sound change: The effect of vowel height and duration on the development of distinctive nasalization. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V* (pp. 52–69). Cambridge University Press.

Harrington, J., Kleber, F., & Reubold, U. (2008). Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *The Journal of the Acoustical Society of America*, *123*(5), 2825–2835. https://doi.org/10.1121/1.2897042

Horii, Y. (1980). An accelerometric approach to nasality measurement: A preliminary report. *The Cleft Palate Journal*, *17*(3), 254–261.

Liao, S., Hoole, P., Cunha, C., Kunay, E., Cui, A., Shigemori, L. S. B., Kleber, F., Voit, D., Frahm, J., & Harrington, J. (2022). *Nasal Coda Loss in the Chengdu Dialect of Mandarin: Evidence from RT-MRI*. 1347–1351.

Liao, S., Hoole, P., & Harrington, J. (2023). The relationship between vowel change and nasal loss in the chengdu dialect of mandarin Chinese: Evidence from RT-MRI. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th ICPhS* (pp. 1072–1076). Guarant International.

Ohala, J. J., & Busa, M. G. (1995). Nasal loss before voiceless fricatives: A perceptually-based sound change. *Rivista Di Linguistica*, *7*, 125–144.

Rochet, B. L. (1976). The formation and evolution of the French nasal vowels. *The Formation and Evolution of the French Nasal Vowels*. https://doi.org/10.1515/9783111328287

Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America*, *106*(1). https://doi.org/10.1121/1.427069

What does supraglottic articulatory global speed tell us about disfluencies?

Fabrice Hirsch[1], Ivana Didirková[2], Michael Blomgren[3], Sofiane Azzouz[4], Fanny Guitard-Ivent[1], Slim Ouni[4]

[1]UMR 5267 Praxiling – University Paul-Valéry Montpellier 3 & CNRS
[2]UR 1569 TransCrit – University Paris 8 Vincennes
[3]University of Utah, Department of Communication Sciences & Disorders
[4]UMR 7503 LORIA, University of Lorraine, CNRS & INRIA

fabrice.hirsch@univ-montp3.fr, ivana.didirkova@univ-paris8.fr, michael.blomgren@health.utah.edu, sofiane.azzouz@loria.fr, fanny.guitard-ivent@univ-montp3.fr, slim.ouni@loria.fr

Introduction. Stuttering is a speech disorder that affects 1% of the world's population. Although we still do not know precisely what causes developmental stuttering, recent research suggests stuttering has genetic (Domingues & Drayna, 2015) and neurological causes (see Etchell, 2018). Stuttering is best known for the speech disfluencies it causes. These disfluencies generally take the form of sound prolongations, silent blocks, and/or repetitions of sounds, syllables, or words, which have been the subject of many studies aimed at understanding their acoustic and perceptual characteristics. However, few studies have focused on the articulatory movements present during these disfluencies.

Among the earliest work in this area, Zimmerman (1980) analyzed stuttering-like disfluencies (SLD). His work suggested that interarticulator positions occurring in disfluent utterances produced by stuttering speakers differed from those in fluent utterances of ordinarily fluent speakers. Furthermore, aberrant interarticulator positions preceded repetitive movements and static posturing. Shapiro (1980) observes "abnormal muscular activity" at the supraglottic and laryngeal levels, with excessive muscular activity, and poor coordination. A series of studies by McClean and colleagues (McClean, 2000; McClean et al., 2004; McClean & Runyan, 2000) investigated kinematic differences between stuttering and nonstuttering speakers. They noticed modifications in supraglottic articulatory kinematics when the stuttering severity increases, especially an increase in tongue and lower lip velocity in nonsense sentences and a lower tongue velocity in severe stuttering in meaningful sentences. More recently, a series of studies examining the articulatory characteristics of SLD were carried out by Didirková and collaborators. Didirková et al. (2019) noticed that the classification of SLD based on perceptual characteristics does not correspond to their articulatory realization, which was confirmed by Lu et al. (2022). Indeed, comparable gestures are present during blocks, prolongations, and repetitions. Didirková and Hirsch (2019) pursue this research by observing coarticulation during SLD. Their findings show several articulatory behavior configurations in supraglottic articulatory movements. In another study, Didirková et al. (2021) observed that SLD and other disfluencies (OD) produced by stuttering and nonstuttering speakers have common articulatory characteristics. However, SLD and OD produced by stuttering speakers present some particularities, mainly in terms of movement retention and anticipation.

The present research expands on the work carried out by Didirková et al. The aim is to assess the speed of movement of speech articulators during disfluencies. Based on the theory that the articulatory movements present during SLD are linked to uncontrolled speech-motor movements, we hypothesize that the speed of movement of the articulators makes it possible to differentiate SLD from OD.

Methods. Four people who stutter and four age- and gender-matched control subjects participated in this research. The stuttering speaker group included two males and two females—all four individuals presented with severe stuttering as determined by a SLP using SSI-4 (Riley, 2009). Participants were instructed to discuss a "typical day" or their interests, and no disfluency elicitation techniques were used. Articulatory data were acquired using the EMA (Carstens; AG501 3D). The sampling frequency was 250 Hz, and the device's accuracy was 0.3 mm. An audio recording (44.1 kHz, 16 bits, .wav), synchronized to the EMA, was made parallel with the data collection. Ten coils were attached to the following points: (1) Two were placed in the middle of the lower and upper lips; (2) One was used to follow the movements of the mandible; (3) Three sensors were glued on the tongue (one on the tongue tip, one on the tongue body, and one on the tongue

dorsum); (4) One coil was used to obtain the contour of the palate; (5) Two coils were placed behind the ears and one on the forehead to control head movements.
The articulatory and audio data were transcribed (orthographically and phonetically) in Praat (Boersma & Weenink, 2022) and then segmented into words, syllables, and phonemes with the EasyAlign plugin (Goldman, 2011). The alignment was then corrected manually. The disfluencies produced by all the speakers were annotated. The position of each coil was extracted every 4 ms using Visartico software (Ouni *et al.*, 2012). We then calculated the speed of the upper and lower lips, the mandible, the tongue tip, and the tongue body's vertical movements. A total of 1,291 disfluencies were considered. Statistical analyses were conducted using RStudio (RStudioTeam, 2020). Non-parametric tests for mean comparisons were used, with Bonferroni correction for multiple comparisons.

Results. Our analysis suggests that measured average speed was systematically higher in nonstuttering speakers than in stuttering speakers. An in-depth analysis of the disfluencies reveals differences between the SLD and OD of stuttering and nonstuttering speakers. Surprisingly, results show slower supraglottic articulatory movements in OD produced by stuttering speakers than in SLD. Regarding speed depending on the disfluency type, analyses reveal that repetitions had systematically higher than average speed independently of the articulator. For some articulators, silent pauses, unfilled pauses, and prolongations show lower than average speed. A coefficient of variation was calculated for the global speed of each articulator to determine whether there were any inter-group differences. The results show that, for all articulators except the tongue body, the variation is higher in stuttering speakers than in normally fluent speakers. In stuttering speakers, the variation is generally higher in SLD than in OD, except for the upper lip and the tongue body.

Discussion. The current study's results align with other studies suggesting increased speech-motor variability in stuttering speakers. The results also support an explanation of overall decreased articulatory speed in stuttering speakers. We suggest increased instability could lead to more complicated management of the normally fine-grained balance between sequential and overlapping muscle movements during speech production. This complication in accurate motor timing could decrease speech rate as a natural compensatory strategy to preserve fluency. Thus, our research suggests that the speed of articulatory movements is a likely distinguishing factor between stuttering and nonstuttering speakers. However, this study only observed articulatory speed behavior as a global measure. Further research should expand the study on local measures of acceleration and deceleration during SLD, which would provide other insights related to specific articulatory gestures. In other words, measuring speech rate within smaller speech segments will help determine if a specific section of a disfluent gesture is impaired or if the entire gesture is slowed down.

References

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer. [Computer program]. Version 6.0.50, Retrieved 31 March 2019 from http://www.praat.org/

Didirková, I., & Hirsch, F. (2019). A two-case study of coarticulation in stuttered speech. An articulatory approach. *Clinical Linguistics & Phonetics*, 1-19. https://doi.org/10.1080/02699206.2019.1660913

Didirková, I., Le Maguer, S., Hirsch, F., & Gbedahou, D. (2019). Articulatory behaviour during disfluencies in stuttered speech. *The 19th International Congress on Phonetic Sciences*, Melbourne, Australia, 2991-2995.

Didirková, I., Le Maguer, S., & Hirsch, F. (2021). An articulatory study of differences and similarities between stuttered disfluencies and non-pathological disfluencies. *Clinical Linguistics & Phonetics*, 35(3), 201-221. https://doi.org/10.1080/02699206.2020.1752803

Domingues, C. E. F., & Drayna, D. (2015). The genetics of stuttering. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0025

Etchell, A. C., Civier, O., Ballard, K. J., & Sowman, P. F. (2018). A systematic literature review of neuroimaging research on developmental stuttering between 1995 and 2016. *Journal of Fluency Disorders*, 55, 6-45. https://doi.org/10.1016/j.jfludis.2017.03.007

Goldman, J. P. (2011). EasyAlign: An automatic phonetic alignment tool under praat. Proceedings of InterSpeech, September, Firenze, Italy.

Lu, Y., Wiltshire, C. E., Watkins, K. E., Chiew, M., & Goldstein, L. (2022). Characteristics of articulatory gestures in stuttered speech: A case study using real-time magnetic resonance imaging. *Journal of Communication Disorders*, 97, 106213.

McClean, M. D. (2000). Patterns of orofacial movement velocity across variations in speech rate. *Journal of Speech, Language, and Hearing Research: JSLHR*, 43(1), 205-216. https://doi.org/10.1044/jslhr.4301.205

McClean, M. D., & Runyan, C. M. (2000). Variations in the Relative Speeds of Orofacial Structures With Stuttering Severity. *Journal of Speech, Language, and Hearing Research,* 43(6), 1524-1531. https://doi.org/10.1044/jslhr.4306.1524

McClean, M. D., Tasko, S. M., & Runyan, C. M. (2004). Orofacial movements associated with fluent speech in persons who stutter. *Journal of Speech, Language, and Hearing Research: JSLHR*, 47(2), 294-303. https://doi.org/10.1044/1092-4388(2004/024)

Ouni, S., Mangeonjean, L., & Steiner, I. (2012). VisArtico: A visualization tool for articulatory data. *Interspeech 2012*, Portland, OR, USA.

Riley, G.D. (2009). SSI4 Stuttering Severity Instrument Fourth Edition.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Shapiro, A. I. (1980). An electromyographic analysis of the fluent and dysfluent utterances of several types of stutterers. *Journal of Fluency Disorders,* 5(3), 203-231. https://doi.org/10.1016/0094-730X(80)90029-7

Zimmermann, G. (1980). Articulatory behaviors associated with stuttering : A cinefluorographic analysis. *Journal of Speech and Hearing Research,* 23(1), 108-121. https://doi.org/10.1044/jshr.2301.108

Inter-subject variation in tongue shape during vowel production in /b/V/t/ sequence: An rtMRI study using 8 vowels from 74 subjects

Satyadev Badireddi[1], Shreya Shrikant Karkun[2], Prasanta Kumar Ghosh[1]

[1]*Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India*

[2]*B.M.S. College of Engineering, Basavanagudi, Bangalore-560019, India.*

satyadevb@iisc.ac.in, shreyashri025@gmail.com, prasantg@iisc.ac.in

Introduction. Speech production involves intricate and coordinated movements of articulators—the lips, velum, jaw, and tongue—as a subject produces different sounds. The tongue's flexibility and versatility allow for a wide range of configurations within the oral cavity making it the primary orchestrator in shaping sounds during speech production. This paper focuses on inter-subject variability of the tongue contour as seen on the mid-sagittal plane during vowel articulation in /b/V/t/ sequences. The orientation of the tongue within the oral cavity determines the resonance and airflow, thereby shaping the vowels pronounced. It is known that vowels are classified based on tongue height (high, mid, low) and tongue advancement (front, central, back). While there are general principles guiding articulator positioning for specific vowels, variations in tongue shape and positioning persist. Inter-subject differences in vocal-tract morphology (palate shape) have a significant influence on lingual articulation (Serrurier et al. 2019; Lammert, Proctor, and Narayanan 2013). These results suggest considerable variability in tongue alignment among people. However, a large dataset from diverse speakers is required to perform a comprehensive study that takes the morphological differences into account. A quantitative study in this work is carried out to compare tongue contours during vowel production using real-time MRI (rtMRI) video data from 74 speakers. Quantitative metrics are employed to measure consistency in the tongue shape, following a simple morphological normalization among individuals for 8 different vowels (ɔ as in 'bought',ʌ as in 'but', oʊ as in 'boat', uː as in 'boot', ɛ as in 'bet', ɪ as in 'bit', æ as in 'bat', iː as in 'beat') spoken in the context of b and t i.e., /b/V/t/.

Methods. For the experiments, we use USC-TIMIT corpus (Lim et al. 2021) consisting of 2D midsagittal-view rtMRI videos (with synchronized audio) of subjects while they speak given stimuli. Vocal tract image (at the middle of the vowel segment) data of 74 individuals articulating /b/V/t/ sequences has been extracted from the video frame and compiled for analysis. The tongue contours are extracted manually, following steps as outlined in Valliappan, Mannem, and Ghosh (2018) from these images and they form the basis for this study. Tongue contours for eight vowels for a subject are illustrated using green curves in Figs. 1(a)-(h). To standardize these outlines, normalization procedures have been implemented to align the tongue contours from different subjects in relation to two fixed anatomical reference points, which are also marked manually—the nose tip (NT) and the velum's projection onto the pharyngeal wall (VEL-P) (also shown in Fig 1 (a) - (h)), known for their relative stability during speech production. These contours undergo adjustments in rotation and scaling, orienting them such that the line connecting the NT and VEL-P becomes horizontally aligned as shown in Figure 1(i), (where tongue contours averaged across all subjects are shown for every vowel). Every tongue contour is resampled into 100 equidistant points, ensuring uniform spacing and equal segment lengths along the contour. This is done to facilitate uniformity across tongue contours and standardize calculations. Let (x_k^i, y_k^i), for $k = 1 \dots 100$ be the coordinates representing the i-th subject's tongue contour for a vowel. We compute ρ_x^{ij} as the correlation coefficient between two 100-dimensional vectors, $(x_1^i \dots x_{100}^i)$ and $(x_1^j \dots x_{100}^j)$. The same procedure is followed for ρ_y^{ij}. Thus, we get ${}^{74}C_2 = 2701$ correlation coefficient values separately for x and y coordinates for a vowel. The mean of these two sets of 2701 correlation coefficients are denoted by ρ_x and ρ_y, for x and y coordinates, respectively. Then, we carry out the following statistical test with the null hypothesis $H_0 : \rho_x = \rho_0$. Based on the methods described in Hinkle, Wiersma, Jurs, et al. (2003), z-transform is applied on ρ_x^{ij} values for testing. Let us call the resultant values by $z_x^{ij} = \frac{1}{2} \ln(\frac{1+\rho_x^{ij}}{1-\rho_x^{ij}})$. ρ_0 is varied from 0 to 1 and significance test is conducted on the data samples z_x^{ij} using z-transformed values of ρ_0. The p-value from the statistical test is used as a measure to indicate the value of ρ_0 for which the null hypothesis can not be rejected indicating the degree of consistency among tongue contours, on average. This is also repeated for ρ_y.

Results. As ρ_0 is varied from 0 to 1, for the null hypothesis $H_0 : \rho_x = \rho_0$, the p-value is calculated and plotted on the y-axis with ρ_0 on the x-axis in Fig.1(j). However, the p-value < 0.05 for the entire [0,1] range except for specific values of ρ_0 within 0.95 and 1.00. In this range [0.95,1], ρ_0 is varied in a step of 0.0005 and the p-value is calculated. These specific

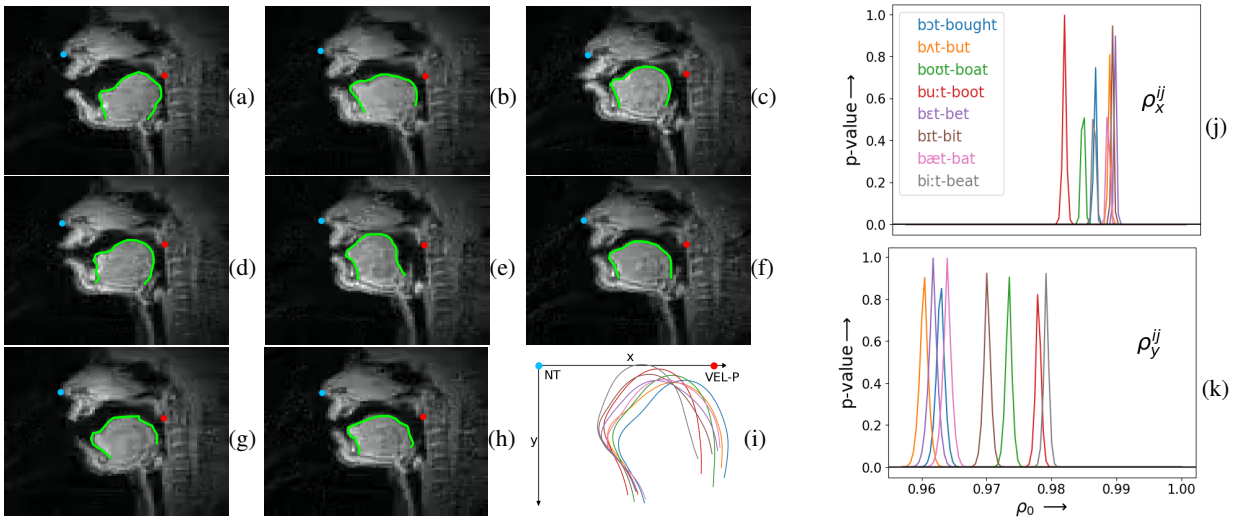


Figure 1: (a)-(h) - Tongue Contour, NT and VEL-P for vowels (/ɔ/ ,/ʌ/, /oʊ/, /u:/, /ɛ/, /ɪ/, /æ/, /i:/), (i) - Normalized contour, averaged across all subjects, (j)-(k) - p-value vs ρ_0 for ρ_x^{ij} and ρ_y^{ij} respectively

values change depending on the vowel as indicated by different colors in Fig. 1(j). For example, for $0.98 < \rho_0 < 0.983$, the H_0 can not be rejected for vowel /u:/ and for $0.987 < \rho_0 < 0.99$, the H_0 can not be rejected for vowel /ɛ/. All the other vowels are between 0.98 and 0.99 as clear from the peaks in the plot in Fig 1(j). Similarly, Fig.1(k) shows p-value vs ρ_0 for $H_0 : \rho_y = \rho_0$. It is clear that for $0.958 < \rho_0 < 0.962$, the H_0 can not be rejected for vowel /ʌ/, and for $0.977 < \rho_0 < 0.981$, the H_0 can not be rejected for vowel /i:/. The rest of the vowels are between 0.958 and 0.981.
The shape of the tongue is determined by the sequence of its X and Y coordinates. Results in Fig.1 indicate that, on average, the tongue shapes for a vowel from any two subjects are highly correlated with a correlation coefficient more than 0.95. These results suggest that with simple morphological normalization, the tongue shape appears to be consistent across 74 subjects in this study, although each of these subjects may have varied vowel-specific articulatory targets influenced by their respective morphology.

Discussion. Vocal tract morphology changes from subject to subject. Thus, for achieving an acoustic target every subject may perform their morphology-specific alteration in articulatory target. This altered articulatory target may appear to be different for different subjects even though they may correspond to the same acoustic target, e.g., vowels in this study. However, the outcome of this study suggests that with two fixed points based simple morphological normalization, the articulatory targets were found to be consistent for each of the eight vowels, considered in this study. Better morphological normalization techniques may remove subject-specific variabilities in tongue shape even more accurately revealing more consistency among tongue shapes from different subjects. A perfect morphological normalization would aim for identical tongue shapes across subjects leading to the metric used in this study $\rho_x = \rho_y = 1$. However, we did not achieve that probably due to several reasons including the simple normalization technique used. It is also interesting to note that tongue shapes for different vowels reveal different degrees of consistency as observed from the peaks in the plots in the Figs.1(j)-(k). Also, this order (in terms of the degree of consistency across different vowels) is not the same for ρ_x^{ij} and ρ_y^{ij}. This could be because tongue height is controlled more by the values in the Y direction and less in the X direction. This also suggests that we need to examine better ways of representing the tongue shape and study consistency using that representation. These are parts of our future works.

References.

Hinkle, Dennis E, William Wiersma, Stephen G Jurs, et al. (2003). *Applied statistics for the behavioral sciences*. Vol. 663. Houghton Mifflin Boston.

Lammert, Adam, Michael Proctor, and Shrikanth Narayanan (2013). "Interspeaker variability in hard palate morphology and vowel production". In.

Lim, Yongwan, Asterios Toutios, Yannick Bliesener, Ye Tian, Sajan Goud Lingala, Colin Vaz, Tanner Sorensen, Miran Oh, Sarah Harper, Weiyi Chen, et al. (2021). "A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images". In: *Scientific data* 8.1, p. 187.

Serrurier, Antoine, Pierre Badin, Laurent Lamalle, and Christiane Neuschaefer-Rube (2019). "Characterization of inter-speaker articulatory variability: A two-level multi-speaker modelling approach based on MRI data". In: *The Journal of the Acoustical Society of America* 145.4, pp. 2149–2170.

Valliappan, CA, Renuka Mannem, and Prasanta Kumar Ghosh (2018). "Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video Using Semantic Segmentation with Fully Convolutional Networks". In: *Proc. Interspeech 2018*, pp. 3132–3136. DOI: 10.21437/Interspeech.2018-1939.

A pulse-step model of speech motor control: Evidence for an extrinsic pacemaker

Alan A. Wrench[1, 2]

[1]Queen Margaret University, UK
[2]Articulate Instruments Ltd
awrench@articulateinstruments.com

Introduction. Most models of movement, including speech production, are predicated on a continuous descending motor command signal. In previous work we have described a discrete pulse step model of movement control which is compatible with biological observations. Recordings by Fetz (2002) of motor cortex neurons show 76% exhibit a pulse-step form during a wrist flexion movement. A very simple limb model (Wrench & Balch, 2017) demonstrates how movement can be modelled by an initial synchronous pulse of activation signals to all the muscles involved in the movement followed by a step change in activation levels halfway through the planned movement. The initial pulse sets nominal muscle contraction lengths such that, given time, the movement end-effector would settle at a point beyond the target. The further beyond the target, the greater the velocity of the initial movement. The step change in activation, halfway through the movement, adjusts the activation levels so the nominal muscle lengths cause the end-effector to come to rest at the target position. This in effect acts as a brake on the movement. Thus, the initial pulse defines the velocity and direction of a movement while the step brakes the movement. Contracting all the muscles simultaneously usually causes significant jerk. To minimize the jerk, which can strain muscles, we propose that inhibitory interneurons in the motor nucleus, delay the activation of the involved muscles. We demonstrated previously, using our simple limb model, how optimal delays to the various agonist and antagonist muscles produce a bell-shaped velocity profile, straighten the trajectory (of limb movement), and match observed staggered EMG pulse patterns seen in real-life movements that the model is based on. It is notable that this discrete model of movement control explains observed EMG activation patterns.

Fluent sequential movements can be handled slightly differently. When starting to learn a movement from A to B to C it would be broken into two transitions with two velocity peaks (initiate-brake-initiate-brake). However, *after practice* Sosnik et al (2004) show that a new more efficient movement plan can be learnt with a single velocity peak. To achieve this in our model, the initial velocity and direction is altered and the step activation is set to land on target C but passing through target B as it does so. We propose that for speech, such efficient plans are practiced, memorized, and recalled as an articulatory ABC phone sequence plan to replace an AB phone transition plan followed by BC plan.

The movement described so far requires proprioceptive feedback from muscle spindle afferents but is primarily a feedforward process based on expected and learnt mappings between muscle activations and positions in sensory space. But muscles can tire, and unexpected external forces may be applied. Visual, auditory and somatosensory feedback of any difference between the expected and actual movement may act to adjust muscle activation levels and correct the movement. The Cerebellum is known to map descending muscle activation levels to expected sensory inputs (Fautrelle et al , 2011). In our model a corrective error signal would be applied at the next available pulse or step. Over time, if the new conditions persist, the cerebellar errors diminish, the muscle activation-sensory input map is thus redrawn and the pulse step plan adapts to the new map.

An important question arises from this discrete model. What determines the timing of the pulses and steps, particularly in a fluent sequence of movements such as speech? We hypothesis that the movements are paced and gated by one or more synchronised extrinsic clocks and look for evidence.

Methods. We employ pose estimation of tongue keypoints from ultrasound (Wrench & Balch-Tomes, 2022) to generate five distance measures from short tendon to points along the tongue body corresponding to independently controlled sectors of the genioglossus muscle. We also record vertical lip separation from pose estimation of video. The measurements are calculated and displayed in the Articulate Assistant Advanced software (Articulate Instruments Ltd, 2023). The periodicity of the vertical lines is manually adjusted until they align with beginnings and ends of articulatory movement transitions corresponding to five genioglossus compartments (charted in green, red, blue, purple, & pink). These distance measure from short tendon to points on the tongue surface are also displayed in 2D midsagittal space (top right, figure 1). Word pairs differing in syllabic stress have been recorded from a single adult speaker. The spoken utterance in figure 1 is the last part of the sentence "The Presbyterian minister managed to curb the drin**king habits of the loitering youths**." by a 60+ year old male subject.

Results. Results are inconclusive. While some recordings show that evenly paced time points can be fitted so that they correspond to initiation of genioglossus muscle compartment length changes throughout an utterance, in others, this simple regular pattern cannot be fitted. The period between solid lines in Figure 1 is 105ms but in other adult utterances where regular timing can be seen, we have observed the period to vary.

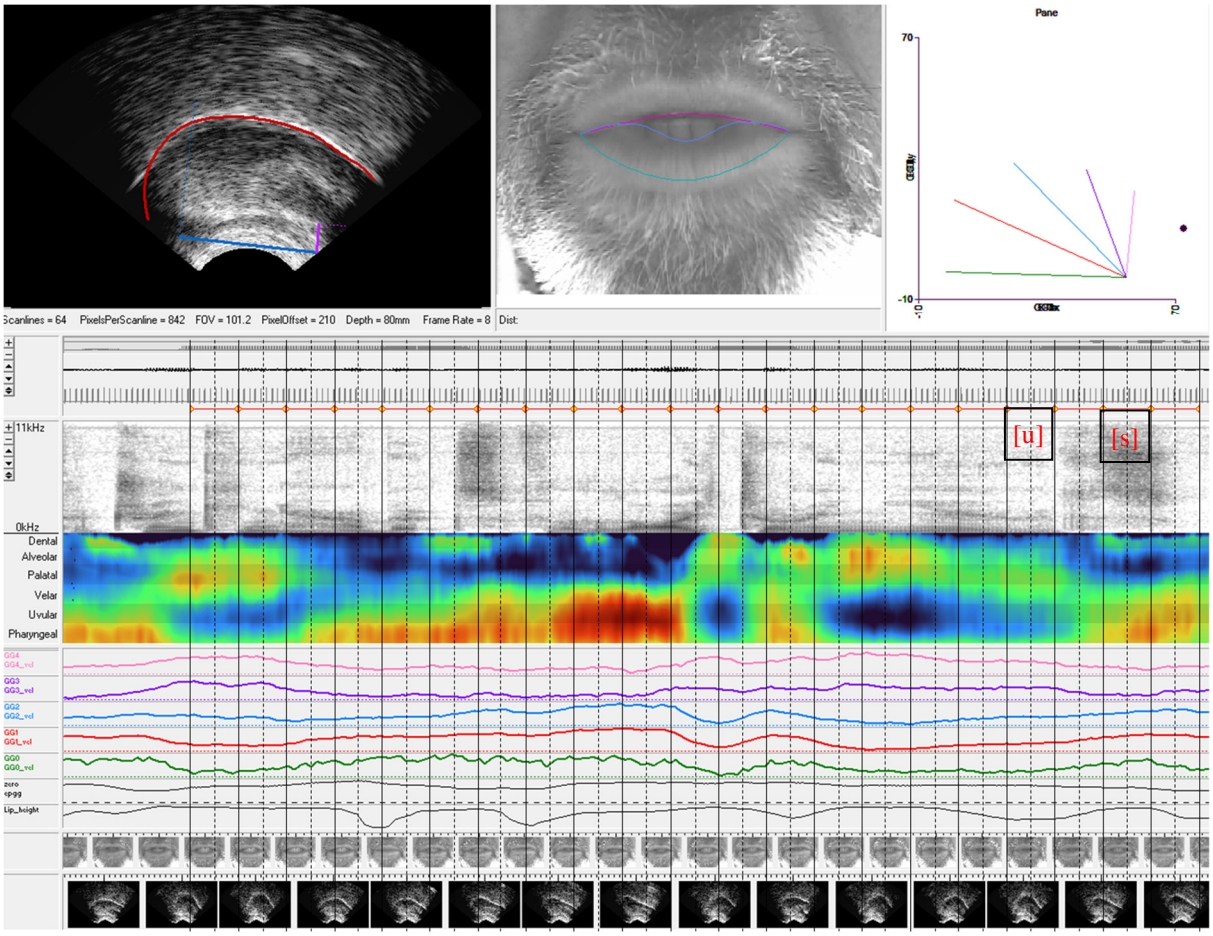


Figure 1: *Shows tongue contour, short tendon and hyoid positions in ultrasound (81Hz) and lip contours from video (60Hz) using DeepLabCut pose estimation. Coloured charts show variation in lengths of different compartments of genioglossus. Vertical lip separation is shown in bottom trace. Glossogram (below spectrogram) shows vocal tract constrictions (Orange-Red) over time.*

Discussion. Results are inconclusive. Regular pacing can be fitted to some recordings such as Figure 1. In other recordings this is not possible. Where this pacing holds we see that prosody sits on top of this pacemaker rhythm. For much of the utterance, muscular activations reach their target in a single pacing period and then one or more muscles immediately transition to a new target. However, prominence and phrase final lengthening appears to be achieved either by sustaining the muscle target for an extra cycle or reducing the velocity so the transition takes longer to reach the target. This can be seen in Figure 1 in both the final vowel [u] and final consonant [s] of "youths". We continue to make improvements to instrumentation and to further investigate this theory.

References

Fetz, E.E., Perlmutter, S, Prut, I.Y., Seki, K., Votaw, S., (2002) Roles of primate spinal interneurons in preparation and execution of voluntary hand movement, Brain Research Reviews 40 pp53–65

Fautrelle, L. Pichat, C., Ricolfi, F., Peyrin, C. ,Bonnetblanc F., (2011), Catching falling objects: the role of the cerebellum in processing sensory–motor errors that may influence updating of feedforward commands. An fMRI study, Neuroscience, Volume 190, Pages 135-144, ISSN 0306-4522, https://doi.org/10.1016/j.neuroscience.2011.06.034.

Sosnik, R., Hauptmann, B., Karni, A. Flash, T., (2004) When practice leads to co-articulation: the evolution of geometrically defined movement primitives. Exp Brain Res 156, pp422–438. https://doi.org/10.1007/s00221-003-1799-4

Wrench, A & Balch, P (2017) A massless 3D biomechanical model of the tongue and its relation to the λ model, 7th International Conference on Speech Motor Control Groningen: Abstracts, Jaargang 22, Supplement, pp 10

Wrench, A., and Balch-Tomes, J., 2022. Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut" Sensors 22, no. 3: 1133. https://doi.org/10.3390/s22031133

Auditory Feedback Perturbation of F2 in French-speaking Children

Isabelle Démosthènes[1,2], Lucie Ménard[1,2]

[1] Université du Québec à Montréal
[2] Centre de recherche sur le cerveau, le langage et la musique
demosthenes.isabelle@courrier.uqam.ca, menard.lucie@uqam.ca

Introduction. Auditory feedback perturbation is known as an efficient tool to understand the role of auditory feedback in speech production and has been studied by many in the past decades (Caudrelier & Rochet-Capellan, 2019). In typically hearing children, acoustic feedback plays a crucial role in guiding their construction of a model to support speech fluency. Without adjustments in articulation, alterations of the shape, size, and strength of speech articulators could profoundly affect acoustic outputs (Callan et al., 2000; Guenther, 1994). Once this model matures, the feedforward system takes over the control of articulators. Despite the critical role of auditory feedback in speech development, only a handful of studies have investigated the impact of sensory manipulation on speech motor control in children. Yet, Littlejohn and Maas (2023) suggest that tasks like feedback perturbation could help researchers and clinicians to better identify and understand the breakdowns in different speech impairments and help differential diagnosis. But to do so, a complete understanding of the processes involved during development is needed. In this context, our project follows the work of Trudeau-Fisette et al. (in review) and aims to pursue the investigation of the development of sensorimotor relationships through compensatory responses to real-time auditory feedback perturbations by comparing adult performance to that of non-reading preschool children. Where the latter focused on the labiality contrast, we will be investigating the place of articulation phonetic feature, implemented along the F2 dimension, and traditionally known to be related to front-back tongue dimension only.

Methods. 15 adults (age 18-35) and 15 children (age 4-6) with no known neurodevelopmental disorder will be screened for hearing impairment and presented with two tasks. First, an auditory identification task, displayed using PsychoPy (v2022.2.4) will invite participants to select the vowel they perceive between /o/ and /ø/ by clicking on the corresponding picture ("*eau"* /o/, water or "*eux"* /ø/, them). Each stimulus of the 10-step continuum built using the Maeda model (Maeda, 1979) will be randomly presented seven times. Then, in a real-time auditory perturbation task using Audapter (Cai et al., 2008), productions of the vowel /ø/ will gradually be shifted toward /o/ by lowering F2 up to 30% through five phases: reference (no shift, four repetitions of six target words with the structure /pV/ giving reference productions for /i, u, a, o, ø, y/), baseline (no shift, 10 utterances of /ø/), ramp (1% decrease shift per trial, 30 utterances of /ø/), hold (30% shift, 15 utterances of /ø/), end (no shift, 15 utterances of /ø/). To ensure that participants hear only their production through the system, a white noise will be presented in the headphones throughout the perturbation task. To avoid a Lombard effect or discomfort for the participants, a good signal-to-noise ratio will be ensured with the microphone's gain. Identification task data will be analyzed in Matlab (R2022b, Update 7) using the Probit regression method to extract the slope of the labeling function and the 50% crossover category boundary. For the feedback perturbation task, mean F1, F2 and F3 values will be extracted in Praat (v. 6.1.16) in the time interval 20 ms before and after midpoint for each vowel. To allow for intersubject comparison, the frequency obtained for each trial will then be normalized using the following formula: $\frac{trial's\ mean\ formant\ value\ (Hz)}{subject's\ mean\ formant\ value\ during\ baseline\ (Hz)}$. A ratio around 1 indicates no changes in production. Values above 1 show an increase in frequency compared to baseline (opposite to the perturbation applied for F2) whereas values below 1 indicate a decrease (following the perturbation applied for F2). Like Trudeau-Fisette et al. (in review), we will use a linear mixed effects model to investigate the effect of the group (Adult vs Children), the experimental phase (Baseline, Ramp, Hold, End) and the trial number (first three trials and last three trials) on the ratio. We will also look into the relationship between the performance at the perceptual identification task and the baseline F2 variability on the normalized ratios during the hold phase for both groups with multiple linear regression.

Results. Normalized frequencies during experimental trials for five adults and four children are presented in **Figure 1**. Despite the limited data in our preliminary results, we have found some variability between individuals as documented in previous studies (Caudrelier & Rochet-Capellan, 2019). Some participants clearly showed a compensatory response whereas others, compensated less or even followed the perturbation. As expected considering the amplitude of our perturbation (Katseff et al., 2012), we also observed an incomplete compensation for the perturbation. In addition, although our perturbation affected F2, some participants did modify their F1 to compensate for the perceived discrepancy

suggesting that, even though Klein et al. (2019) found no consistent effect of F2 shift on F1, we should consider the relation between formants in our analysis.

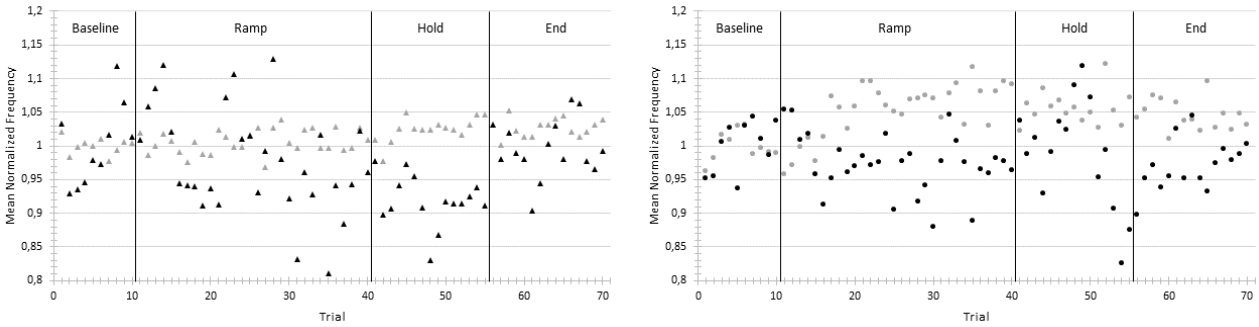


Figure 1: *Mean normalized formant values by phase for /ø/ in our preliminary data. Triangles on the left refer to F1 and dots on the right to F2 for 5 adults (grey) and 4 children (black).*

Discussion. Based on MacDonald et al. (2012) and Trudeau-Fisette et al. (in review) we expect a compensatory response in both groups with a greater variability in children. Like Trudeau-Fisette et al. (in review), we also expect different compensatory profiles in children and adults with more following responses preceding the compensation in the ramp phase, and more following responses overall in children. Many factors have been found to affect the magnitude of the response including vowel used, phonemic structure of the language, shift proximity to perceptual boundary, alteration degree, one's perceptual acuity, within speaker trial-to-trial variability (Caudrelier & Rochet-Capellan, 2019; Trudeau-Fisette et al., in review). When exploring the direction of the perturbation in the F2 dimension in Russian-speaking adults, Klein et al. (2019) did find a response to perturbations for both increase and decrease of F2 albeit having a smaller compensation for downward shifts in some participants. Comparing our results with those of Trudeau-Fisette et al. (in review), will allow us to see if we can also find a response to perturbation in both directions along the F2 axis in French and in children. Furthermore, Trudeau-Fisette et al. (in review) have found differences between children and adults in the factors predicting the amount of compensation observed in the holding phase: in adults, only identification task's labeling function slope had an effect as opposed to F2 variability during baseline being the only that had effects on children. It will be interesting to see if our results follow the same trend.

References

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. F. (2008). A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/. *Proceedings of the 8th Intl. Seminar on Speech Production*, 65–68.

Callan, D. E., Kent, R. D., Guenther, F. H., & Vorperian, H. K. (2000). An Auditory-Feedback-Based Neural Network Model of Speech Production That Is Robust to Developmental Changes in the Size and Shape of the Articulatory System. *Journal of Speech, Language, and Hearing Research*, *43*(3), 721–736. https://doi.org/10.1044/jslhr.4303.721

Caudrelier, T., & Rochet-Capellan, A. (2019). Changes in speech production in response to formant perturbations: An overview of two decades of research. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (Eds.), *Speech production and perception: Learning and memory* (Vol. 6, pp. 15–76). Peter Lang.

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, *72*(1), 43–53. https://doi.org/10.1007/BF00206237

Katseff, S., Houde, J., & Johnson, K. (2012). Partial Compensation for Altered Auditory Feedback: A Tradeoff with Somatosensory Feedback? *Language and Speech*, *55*(2), 295–308. https://doi.org/10.1177/0023830911417802

Klein, E., Brunner, J., & Hoole, P. (2019). Spatial and temporal variability of corrective speech movements as revealed by vowel formants during sensorimotor learning. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (Eds.), *Speech production and perception:Learning and memory* (pp. 77–107). Peter Lang.

Littlejohn, M., & Maas, E. (2023). How to cut the pie is no piece of cake: Toward a process-oriented approach to assessment and diagnosis of speech sound disorders. In *International Journal of Language and Communication Disorders*. John Wiley and Sons Inc. https://doi.org/10.1111/1460-6984.12934

Maeda, S. (1979). An articulatory model of the tongue based on a statistical analysis. *The Journal of the Acoustical Society of America*, *65*(S1), S22–S22. https://doi.org/10.1121/1.2017158

MacDonald, E. N., Johnson, E. K., Forsythe, J., Plante, P., & Munhall, K. G. (2012). Children's development of self-regulation in speech production. *Current Biology*, *22*(2), 113–117. https://doi.org/10.1016/j.cub.2011.11.052

Trudeau-Fisette, P., Vidou, C., & Ménard, L. (in review). *The development of sensorimotor relationships in speech: Adaptation to real-time auditory feedback perturbations*.

The Role of Executive Functions and Levodopa on Articulatory Timing

Elisa Herbig[1], Tabea Thies[1,2], Michael T. Barbe[2], Doris Mücke[1]

[1] IfL Phonetics, University of Cologne, Germany
[2] Department of Neurology, University Hospital Cologne, Germany
{eherbig1; tabea.thies; doris.muecke}@uni-koeln.de, michael.barbe@uk-koeln.de

Introduction. Speech production requires the control over cognitive functions and motor processes, both of which are affected in Parkinson's disease (PD). While gross motor symptoms like bradykinesia, rigidity, and resting tremor are prominent, the impact extends to speech impairment, characterized by hypokinetic dysarthria, and cognitive dysfunctions. PD-related speech impairment is linked to a hypo-functioning speech system and reduced fine motor control. The deficiencies in speech motor control not only hinder the preparation and maintenance of motor programs but also impede the ability to switch between them (Spencer & Rogers 2005). PD also affects cognitive processes, including working memory, attention, executive control, and visuospatial domains. While executive functions play a crucial role in orchestrating cognitive processes, cognitive efficiency refers to the ability of an individual's cognitive processes to perform tasks with minimal effort while maximizing accuracy and effectiveness. Processing speed and task-switching abilities are key components of cognitive efficiency, which can both be assessed with the Trail Making Test (TMT). In the present study, we investigate the interplay between speech motor control and cognitive dysfunction by examining kinematics of syllable coordination patterns of the complex consonant cluster /pl/ in people with PD (PwPD) and healthy controls (HC). We also explore the role of levodopa on these timing patterns.

Methods. 22 HC (19 m, 3 f, mean age = 60 years) and 25 PwPD (20 m, 5 f, mean age = 60 years) participated in the study. PwPD were diagnosed with PD 7 ± 4 years prior to study inclusion and were recorded in both OFF and ON medication condition. The OFF condition involved withdrawing PD medication for at least 12 hours, while the ON condition entailed the intake of a predetermined supramaximal levodopa dosage of 200 mg. Speech data were recorded acoustically and kinematically using 3D electromagnetic articulography (EMA, AG501). The speech material consisted of words with simple and complex onsets with initial syllables of the target words following either CV (either C_1V /pina/ or C_2V /lina/) or CCV structure (C_1C_2V /plina/). Participants were instructed to embed the target words in a predefined sentence ("Er hat wieder … gesagt" | "He said … again") and to produce it twice. To analyze articulatory timing patterns of the initial consonant clusters, EMA sensors were placed on the lower lip, tongue tip and tongue body. Speech data were processed in the EMU-webApp. On the acoustic level, we calculated segment durations. On the kinematic level, target positions of the articulators for consonants (C) and vowels (V) in the first stressed syllables were identified in the vertical plane using zero-crossings in the respective velocity trace. Latencies between the maximum target positions of the leftmost C (LMC) and the rightmost C (RMC) to the V were computed. All parameters are compared between syllables with low and high complexities, CV and CCV.
We used the C-center coordination paradigm for the kinematic analysis: When a C is added to a CV syllable to form a complex CCV onset, the coordination of Cs and Vs is reorganized. This can be measured in terms of articulatory overlap patterns (Pouplier 2012). To analyze the syllable coordination, we used the following overlap measures: The *leftward shift* is captured by comparing the latency from C_1 to V in the syllable C_1V (/pi/) with C_1C_2V (/pli/) (latency should increase from CV to CCV). The *rightward shift* is usually present from C_2V (/li/) to C_1C_2V (/pli/) (latency from C_2 to V should decrease from CV to CCV). The *C-center* was determined as midpoint between both Cs in CCV syllables and its latency to the vocalic target was calculated. Note that, due to biomechanical shortening, the rightward shift of C_2 can be blocked in /pl/ in German, i.e., due to coarticulatory effects of the jaw, lips, and tongue, the duration of C_2 can be modified systematically (Mücke et al. 2020). By using linear mixed effect models, articulatory timing patterns were compared between groups and medication conditions: (i) HC vs. OFF, (ii) HC vs. ON, and (iii) OFF vs. ON. In addition, participants' executive functions were assessed by using the TMT. PwPD completed the TMT in medication-ON condition. To examine if articulatory timing patterns are dependent on executive functions, phonetic parameters will be correlated with the time participants needed to complete the task. Cognitive efficiency was assessed by dividing the TMT score of part B (measure of executive functions/set-shifting) by the TMT score of part A (measure of processing speed).

Results. Results show that acoustic segment durations of /p/ (C_1) and /i/ (V) do not differ between syllable structures (Table 1). However, acoustic durations of /l/ (C_2) are shorter in CCV syllables compared to CV syllables ($p < .001$ across all comparisons). On the kinematic level, latencies between C_1 (/p/) and the vowel /i/ increase from CV to CCV ($p < .001$ across all comparisons, Figure 1, left) but the latencies between the /l/ (C_2) and /i/ do not change dependent on syllable structure (Figure 1, middle). When comparing the groups (HC vs. OFF), the PwPD in the OFF condition present with longer C_2 ($p = .018$) and V durations ($p = .019$) as well as larger latencies between the RMC and the vocalic anchor ($p =$

.027, Figure 1, middle). V durations ($p < .001$) and latencies between the C targets and the vocalic anchor decrease from OFF to ON condition (LMC: $p = .003$, RMC: $p = .003$). The OFF/ON effect is further reflected in the shortening of the latency between the C-center and the vocalic anchor ($p = .004$). In addition, durations and latencies decrease from OFF to ON to a similar level as HC, eliminating group differences. Moreover, the performance time of the TMT did not differ between the groups, neither of part A nor part B. However, the values of part B (indicator of set-shifting) and the difference ratio TMTB/TMTA (indicator of cognitive efficiency, Figure 1, right) correlate with the latency difference of the RMC (part B: $r = .43$, $p = .003$, ratio: $r = .54$, $p < .001$) in the PD group. Such correlations are not found in the HC group.

Table 1: *Means and standard deviations of relevant measures.*

		Acoustic segment durations (ms)			Articulatory latencies (ms)		
		/p/	/l/	/i/	LMC to V /p/ → /i/	RMC to V /l/ → /i/	C-center → /i/
HC	C_1V	203 (73)	-	121 (39)	185 (80)	-	-
	C_2V		99 (38)	137 (41)		122 (55)	-
	C_1C_2V	205 (78)	55 (21)	115 (33)	248 (71)	115 (43)	182 (48)
OFF	C_1V	199 (44)	-	145 (33)	199 (47)	-	-
	C_2V		131 (59)	163 (42)		150 (45)	-
	C_1C_2V	197 (72)	66 (23)	139 (50)	271 (65)	143 (49)	207 (52)
ON	C_1V	186 (46)	-	131 (34)	187 (52)	-	-
	C_2V		123 (70)	141 (34)		137 (41)	-
	C_1C_2V	189 (66)	53 (18)	121 (29)	243 (55)	128 (37)	186 (43)

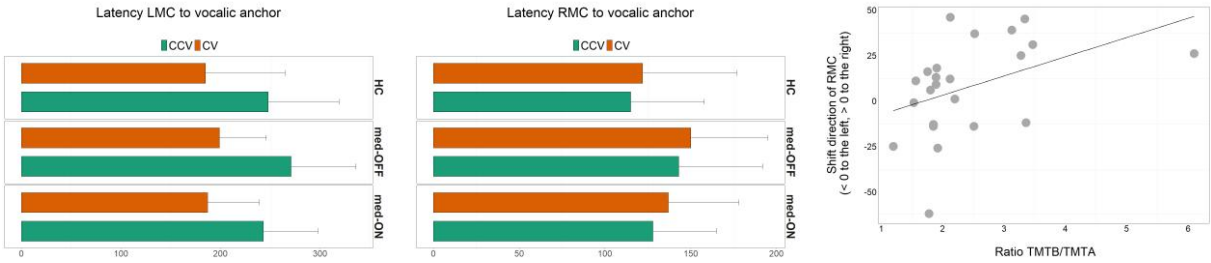


Figure 1: *Latencies between the leftmost C (LMC) and the rightmost C (RMC) comparing CV and CCV syllables, and correlation between RMC shift and cognitive efficiency (TMTB/TMTA ratio).*

Discussion. As expected, the kinematic results reveal a non-symmetrical timing pattern for the complex onset coordination /pl/ for neurotypical speakers of German. The rightmost C does not shift towards the following V from CV to CCV, but the C_2 segment was shortened in CCV (e.g. Pouplier 2012, Mücke et al. 2020). The same non-symmetrical timing pattern was preserved by PwPD for complex syllable organization, even in a poor motor status, i.e. without medication. Our results on inter-gestural timing patterns extend the findings of studies reporting stable and preserved timing patterns in PwPD for vowel productions (e.g. Yunusova et al. 2008). However, we found an effect of levodopa on durations of Cs and Vs: In the OFF condition, PwPD produced longer consonantal and vocalic movements on the intra-gestural level, and the durational changes led to longer latencies between Cs and Vs on the inter-gestural level. This underlines a beneficial effect of levodopa on speech planning abilities, which has been shown before (e.g. Thies et al. 2021). The relationship between timing patterns and cognitive skills demonstrates that PwPD produce a deviant pattern for the RMC shifts as a reflex of the complex onset parse for /pl/ in German. Specifically, the RMC tends to shift to the right when there is a decline in set-shifting abilities, indicating a less efficient timing.

References

Mücke, D., Hermes, A. & S. Tilsen (2020). Incongruencies between phonological theory and phonetic measurement. *Phonology* 37(1), 133-170. DOI: https://doi.org/10.1017/S0952675720000068

Pouplier, M. 2012. *The gestural approach to syllable structure: Universal, language-and cluster-specific aspects.* In Fuchs, Susanne, Weirich, Melanie, Pape, Daniel & Perrier, Pascal (eds.) Speech planning and dynamics. Frankfurt am Main: Peter Lang. 63–96.

Spencer, K. A., & Rogers, M. A. (2005). Speech motor programming in hypokinetic and ataxic dysarthria. *Brain and Language, 94*(3), 347–366. DOI: 10.1016/j.bandl.2005.01.008

Thies, T., Mücke, D., Dano, R., & Barbe, M. T. (2021). Levodopa-based changes on vocalic speech movements during prosodic prominence marking. *Brain Sciences, 11*(5), 594. DOI: 10.3390/brainsci11050594

Yunusova, Y., Weismer, G., Westbury, J. R., & Lindstrom, M. J. (2008). Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research, 51*(3), 596 – 611. DOI: https://doi.org/10.1044/1092-4388(2008/04

Effects of Pharyngealization and Labialization on Formants in Tashlhiyt

Philipp Buech, Anne Hermes, Rachid Ridouane

Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle)

{philipp.buech, anne.hermes, rachid.ridouane}@sorbonne-nouvelle.fr

Introduction. Pharyngealization and labialization are secondary articulations formed by minor constrictions that accompany other stronger constrictions of the primary category. A certain asymmetry regarding their distribution and their state of investigation can be observed in the literature: pharyngealization occurs in approx. 1% of the languages, while labialization is the most common secondary articulation in the world's languages (Buech, Hermes, & Ridouane, 2022). However, the phonetics of pharyngealization is well investigated, especially in Arabic (e.g., Al-Tamimi, 2017), while the phonetic literature on labialization is sparse. Acoustically, secondary articulations are primarily signaled in the vowels surrounding the consonant. Although pharyngealization and labialization are associated with two different vocal tract modifications, previous work on Moroccan Arabic showed that they both lead to a lowered F2 as the primarily affected acoustic parameter (Zeroual et al., 2011). Because of this similarity, earlier work speculated about an 'equivalence of pharyngealization and labialization' (Jakobson, 2002), but others argue that these two secondary articulations differ in their impact on the formant structure beyond F2, e.g., a higher F1 for pharyngealized consonants, but a lowered F1 for labialized ones (Rose, 1979). We present a study on the effects of these two secondary articulations on the formant structure of adjacent vowels in Tashlhiyt, one of the rare languages to have a contrast of both labialization (for dorsals) and pharyngealization (for coronals) in its phonological system. Previous work on pharyngealization in Tashlhiyt showed that the tongue dorsum is lowered with a maximum difference located at the acoustic offset of the pharyngealized coronals (Buech, Ridouane, & Hermes, 2022). In contrast, the production of labialized dorsals is achieved by more closed and protruded lips at their acoustic mid and a tongue backing at their acoustic offset (Buech et al., 2023).

Methods. Twenty-six speakers of Tashlhiyt (9 females, 17 males) were recruited for this experiment. We constructed CVC logatomes, where the consonant was either a plain (t, d, s, z, l, ɾ) or pharyngealized coronal ($t^ˤ$, $d^ˤ$, $s^ˤ$, $z^ˤ$, $l^ˤ$, $ɾ^ˤ$) and plain (k, g, q, χ, ʁ) or labialized dorsal (k^w, g^w, q^w, $χ^w$, $ʁ^w$). The vowel positions were occupied by the three vowels of Tashlhiyt ([i, a, u]), thus leading to a set of [iCi, aCa, uCu] logatomes for each consonant. The targets were embedded in carrier sentences and repeated three times. A total of 4,909 tokens went into the analysis. We measured F1, F2 and F3 at 90% of the preceding V and 10% of the following V. For formant extraction, we used Praat (Boersma & Weenink, 2023) and adapted the Burg algorithm for the speaker's sex. Afterwards, we converted Hertz values into Bark according to Traunmüller (1990). Bayesian linear mixed models were run on each parameter (F1, F2, F3) and vowel position (preceding V, following V) with by-speaker intercepts and slopes as group-level effects. We used the HDI+ROPE decision rule (Kruschke, 2018) for decision making. We set a uniform ROPE of 0.5 Bark, as this is the half bandwidth from a critical band's center frequency to the edges of its adjacent critical bands (Fastl & Zwicker, 2007). We report the mean and the lower and upper boundaries of 95% of the HDI in cases where we rejected the null value.

Results. Fig. 1 shows the frequency ranges of F2 (x-axis) and F1 (y-axis) for plain vs. pharyngealized, and plain vs. labialized productions by vowel position. In pharyngealized vs. plain segments, we found an effect of pharyngealization on F2 in both vowel positions, but the lowering was slightly stronger in the following ($\hat{\beta}$ = -2.28 [-2.82, -1.69]) than in the preceding V ($\hat{\beta}$ = -2.08 [-2.69, -1.50]). Furthermore, the following V showed also a raised F1 ($\hat{\beta}$ = 0.88 [0.68, 1.07]), but we observed no modification of F3. Pairwise comparisons for each vowel environment revealed that the pattern of a lowered F2 in both vowel positions and an increased F1 in the following V manifested across vowel qualities. Vowel qualities differed in the extent of the formant modification: the strongest modification was found for [i] (preceding V F2: $\hat{\beta}$ = -2.53 [-3.17, -1.88]; following V F2: $\hat{\beta}$ = -2.72 [-3.30, -2.12], following V F1: $\hat{\beta}$ = 1.41 [1.19, 1.63]), followed by [a] (preceding V F2: $\hat{\beta}$ = -2.08 [-2.68, -1.49]; following V F2: $\hat{\beta}$ = -2.28 [-2.83, -1.73], following V F1: $\hat{\beta}$ = 0.88 [0.68, 1.07]) and then [u] (preceding V F2: $\hat{\beta}$ = -1.79 [-2.42, -1.12]; following V F2: $\hat{\beta}$ = -1.70 [-2.27, -1.10], following V F1: $\hat{\beta}$ = 0.72 [0.51, 0.94]). For labialization, we also found a general pattern of a lowered F2 in labialized productions, but no effect on F1 and F3. This F2 drop was slightly stronger in the following V ($\hat{\beta}$ = -3.42 [-4.00, -2.84]) than in the preceding V ($\hat{\beta}$ = -3.33 [-3.91, -2.75]). Comparisons within each vowel context revealed a stronger effect on [i] (preceding V F2: $\hat{\beta}$ = -3.98 [-4.59, -3.40]; following V F2: $\hat{\beta}$ = -3.59 [-4.17, -2.99]) than for [a] (preceding V F2: $\hat{\beta}$ = -3.33 [-3.91, -2.74]; following V F2: $\hat{\beta}$ = -3.42 [-3.97, -2.82]), while the formant structure of [u] was not affected by labialization.

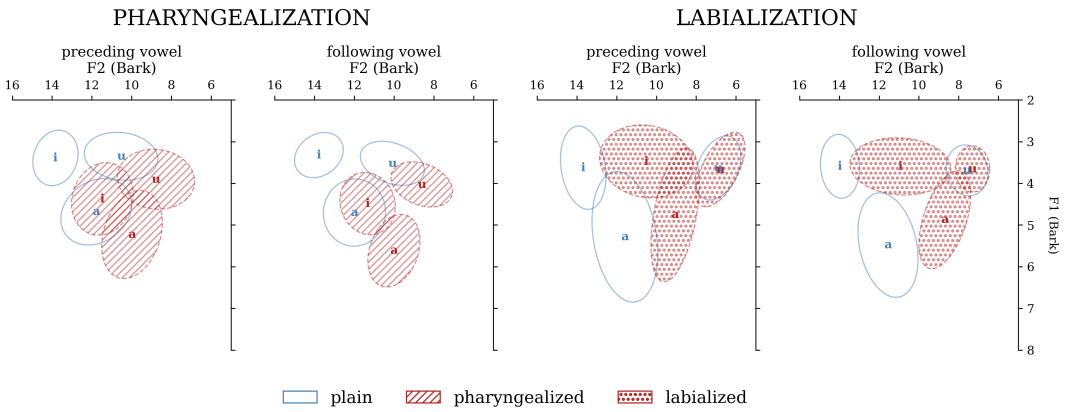


Figure 1: *Formant frequencies of F1 and F2 in Bark by secondary articulation and vowel position. Ellipses show 2.5 standard deviations.*

Discussion. Our results confirm previous studies, indicating a lowered F2 in both pharyngealized and labialized productions. The secondary articulations differed in the extent of F2 modification, where the lowering in labialized productions was stronger than in pharyngealized contexts. Furthermore, they differed in their use of F1, which was raised after pharyngealized coronals but was unaffected by labialization. These acoustic consequences can be explained by the underlying articulation: a retraction of the tongue leads to an F2 lowering and a slight raising of F1, which is enhanced the further back the tongue position is (Lindblom & Sundberg, 1971). While a stronger constriction at the pharynx is a defining attribute of pharyngealization, the tongue backing for labialization is not as strong, and the increase in F1 is not as important. Instead, the characteristic lip protrusion in labialized dorsals results in a lowering of all formants (Lindblom & Sundberg, 1971). This explains why the extent of F2 modification is stronger in labialized compared to pharyngealized productions. The simultaneous actions of tongue backing and lip protrusion enhance the lowering effect of F2 in labialized sounds. In contrast, the effects of tongue backing and lip protrusion on F1 are antagonistic, working together to eliminate any significant influence on this formant. In addition to these differences, we also found similarities: for both pharyngealization and labialization, the F2 lowering was slightly stronger in the following vowel than in the preceding one. The extent to which vowel qualities are affected is also similar, starting with [i] as the vowel with the strongest modification, followed by [a], and [u] being the vowel with the smallest modification (pharyngealization) or no modification (labialization).

References.

Al-Tamimi, J. (2017). Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: Implications for formal representations. *Laboratory Phonology*, *8*(1), 1–40. https://doi.org/10.5334/labphon.19

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer* [Computer program] (6.4.01). Retrieved November 30, 2023, from http://www.praat.org/

Buech, P., Hermes, A., & Ridouane, R. (2022). Towards a typology of secondary articulations. *19èmes rencontres du Réseau Français de Phonologie*. https://doi.org/10.13140/RG.2.2.35426.61126

Buech, P., Hermes, A., & Ridouane, R. (2023). Labialization in Amazigh: acoustic and articulatory marking over time. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th international congress of phonetic science* (pp. 1017–1021). Guarant International.

Buech, P., Ridouane, R., & Hermes, A. (2022). Pharyngealization in Amazigh: Acoustic and articulatory marking over time. *Proc. Interspeech 2022*, 3448–3452. https://doi.org/10.21437/Interspeech.2022-10831

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics. Facts and Models*. Springer.

Jakobson, R. (2002). Mufaxxama – the 'Emphatic' Phonemes in Arabic. In *Volume i phonological studies* (pp. 510–522). De Gruyter Mouton. https://doi.org//10.1515/9783110892499.510

Kruschke, J. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/251524591877130

Lindblom, B. E. F., & Sundberg, J. E. F. (1971). Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement. *The Journal of the Acoustical Society of America*, *50*(4B), 1166–1179. https://doi.org/10.1121/1.1974958

Rose, S. (1979). Phonetic Aspects of Nootka Pharyngeals [Ms.]. https://lingpapers.sites.olt.ubc.ca/files/2018/03/1979%5C_RoseW.pdf

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100. https://doi.org/10.1121/1.399849

Zeroual, C., Esling, J. H., & Hoole, P. (2011). EMA, endoscopic, ultrasound and acoustic study of two secondary articulations in Moroccan Arabic. Labial-velarisation vs. emphasis. In Z. M. Hassan & B. Heselwood (Eds.), *Instrumental studies in arabic phonetics* (pp. 277–297). John Benjamins Publishing Company.

On the Supra-laryngeal Articulation of Prosodic Prominence in Southwestern Mandarin

Jing Huang[1], Feng-fan Hsieh[2], Yueh-chin Chang[3]

National Tsing Hua University

xiaokuidaren94@163.com, ffhsieh@mx.nthu.edu.tw, ycchang@mx.nthu.edu.tw

Introduction. This study is dedicated to exploring the kinematic properties of prosodic prominence in focused vs. non-focused constituents in Southwestern Mandarin (SWM) using Electromagnetic Articulography (EMA). Distinct from Beijing Mandarin, SWM demonstrates a pronounced strong-weak pattern in the acoustic duration of full-toned disyllabic units, as exemplified in studies such as Liu et al. (2022). The aim of this research is to investigate the supra-glottal articulatory dynamics in focused versus unfocused constituents of SWM, an East Asian tone language.

Southwestern Mandarin (SWM), spoken by over 270 million speakers, is the most spoken variety of Mandarin. While phonetically similar to Beijing Mandarin, it lacks neutral tones in content words. Previous studies suggest a trochaic pattern in SWM disyllables (e.g., Liu et al., 2022; Qin, 2015), where the first syllable receives primary stress. However, the precise articulatory differences between stressed and unstressed syllables within a word remain largely unexplored. This study aims to address this gap by investigating the kinemtics of prosodic prominence in SWM, focusing on tongue movement, jaw displacement, consonant plateau duration, (amplitude-normalized) peak velocity, and consonant-vowel lags. We also examine the impact of phrasal prominence, specifically narrow focus, on articulation. To this end, our research questions are as follows:

1. At the word level, do stressed syllables in SWM display longer vowel durations, greater consonant movement amplitude, higher peak velocities, and extended consonant-vowel lags compared to unstressed syllables?
2. What articulatory differences are observable between syllables positioned in focused versus unfocused segments of a phrase?

Methods. The experiment involved the participation of four native SWM speakers, all in their 20s (1 female). Articulatory data was collected using a Carstens AG501 system. The test materials comprised six (possible) personal names: /pe^{33}.pe^{33}/, /pa^{24}.pa^{55}/, /ti^{24}.ti^{55}/, /tu^{24}.tu^{55}/, /tjen^{35}.tjen^{21}/, and /twan^{24}.twan^{55}/ (seven repetitions for each target item). Target words were embedded within the carrier phrase, "*p^{h}e* A_1A_{2FOC}, *pu p^{h}e* B_1B_2", meaning "Pat A_1A_{2FOC}! Do not pat B_1B_2." Contrastive focus is placed on AA (marked as AA_{FOC}), which means the phrasal prominence falls on AA. To investigate the word-level prominence, the target words on the off-focus position (BB) were analyzed to minimize the influence of high-level prominence, such as phrasal prominence, focus, and domain-initial strengthening (Keating et al. 2004). To explore the effect of phrasal prominence, we compared the articulatory properties in A_1 (focused) and B_1 (off-focus). Regarding the articulatory measurements, we used *Mview* (Tiede 2005) to extract the consonant plateau duration and rapidity (=Amplitude and peak velocity; see Roon et al. 2021). The corresponding gestures for each target item are as follows: /pe^{33}.pe^{33}/ (C: LA vs. V: TBz), /pa^{24}.pa^{55}/ (C: LA vs. V: TDz), /ti^{24}.ti^{55}/ (C: TTz vs. TBz), /tu^{24}.tu^{55}/ (C: TTz vs. TDz), /tjen^{35}.tjen^{21}/ (C: TTz vs. TBz), and /twan^{24}.twan^{55}/ (C: TTz vs. TDz). For the quantitative analysis of EMA sensor trajectories, Generalized Additive Mixed Modeling (GAMM) was employed, as detailed in Wood (2016), Wieling (2018), and Sóskuthy (2021).

Results. Regarding Research Question 1, we compared the two identical syllables in B_1 and B_2 positions. Firstly, the jaw movement data show that the final syllables sometimes have more pronounced vertical jaw movement (JAWz), or greater jaw displacement. Secondly, five out of six pairs in B_1 position show more robust vertical and/or horizontal movements found in the Tongue Body (TB) and/or Tongue Dorsum (TD) than in B_2 position. Thirdly, B_1 is significantly longer than B_2 in consonant plateau duration. Likewise, B_1 exhibits significantly higher stiffness compared to B_2. Finally, B_1 and B_2 do not differ in C-V timing differences. In summary, recall that B_1 and B_2 belong to the same word, and they are not on focus in the experimental design. We can say that the differences in B_1 and B_2 basically reflect the word-level prominence, (word stress). The result is consistent with acoustic measurements, according to which the initial syllables are slightly longer, meaning that SWM has trochaic feet (Strong-Weak).

Regarding Research Question 2, we compared syllables in A_1 (focused) and B_1 (off-focus) positions, emphasizing the distinction in emphasis between the two. Firstly, an examination of the jaw movement (JAWz) data from the provided table indicates the absence of significant vertical dimension differences between focused and unfocused elements. Secondly, within A_1 position, only two of the six pairs demonstrate notably more pronounced vertical movements in the Tongue Body (TB) and/or Tongue Dorsum (TD) compared to B_1. Thirdly, it is observed that consonant plateau duration,

amplitude-normalized peak velocity, and C-V lags are more pronounced exclusively in the context of labial onsets. Acoustically, we found the rime duration of A_1 is longer than that of B_1.

Table 1: *A Summary of Phonetic Differences in Focused versus Unfocused Constituents.*

Positions	Jaw displacement	Tongue movement	Consonant Plateau Duration	C-V lag	Rime duration
B_1 vs. B_2 (Both off-focus)	$B_2 > B_1$ (3 out of 6)	TB and/or TD (5 out of 6)	$B_1 > B_2$ (Cor and Lab)	×	$B_1 > B_2$
A_1 vs. B_1 (Focused vs. off-focus)	×	TB and/or TD (2 out of 6)	$A_1 > B_1$ (Lab)	$A_1 > B_1$ (Lab)	$A_1 > B_2$

Discussion. The present study revealed a significant contrast between word-level prominence and "focus prominence." While word-level prominence (B_1 vs. B_2) manifests in hyper-articulated initial syllables with more pronounced tongue movement, longer consonant plateau duration, and rime duration, focus prominence (defined as emphasis on specific words within A_1 positions) doesn't exhibit the same clear articulatory correlates. Unexpectedly, syllables in A_1 positions lack substantial increases in jaw displacement and tongue movement compared to B_1, despite showing statistically significant variations in other potential cues such as C-V lags (specifically with labial stops, though). This divergence from predictions based on English-centric models (e.g., de Jong, 1995) suggests that cross-linguistically speaking, focus realization may not involve hyper-articulation in the same way. Similarly, while our findings are consistent with Erickson and Kawahara (2016) in some respects, they do not provide strong evidence to fully support the claims of the jaw as one of the prosodic articulators. Notably, our data revealed that jaw displacement is occasionally attested in sentence-final positions and more surprisingly, entirely non-significant in focused constituents. This raises intriguing questions about whether the supra-laryngeal articulation plays a more limited role in prosodic/focus prominence for tone languages like SWM, potentially relying more heavily on tonal variations and perhaps some other unknown cues. Future research could explore whether specific subsets of potential articulatory indicators are employed to implement focus in SWM or if alternative explanations, such as tonal adjustments, can account for the observed patterns.

References

De Jong, Kenneth J. (1995). The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation. J Acoust Soc Am. 97(1): 491-504. doi: 10.1121/1.412275.

Erickson, D. & Kawahara, S. 2016. Articulatory correlates of metrical structure: Studying jaw displacement patterns. *Linguistics Vanguard* 2.1: 20150025. doi: doi.org/10.1515/lingvan-2015-0025

Katsika, A. & Tsai, K. (2021). The supralaryngeal articulation of stress and accent in Greek. *Journal of Phonetics*, 88: 101085.

Keating P, Cho T, Fougeron C, & Hsu C. (2004). Domain-initial articulatory strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds), Phonetic Interpretation: Papers in Laboratory Phonology VI. Cambridge, UK: Cambridge Univ. Press. 143–61.

Liu, C., Li, K. & Nolan, F. (2022). Lun shengdiaoyuyande jiezou yu zhongyin moshi [On the speech rhythm and stress pattern of tone languages]. *Journal of Sichuan University (Philosophy and Social Science Edition)*, 240(3), 151–163.

Qin, Z. (2015). Chengduhua de Liandubiandiao yu Yunlüjiegou [Tone sandhi and prosodic structure in Chengdu dialect]. *Hanyu Xuebao [Chinese Linguistics]*, 50(2), 36–44.

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, 84, 101017.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, 70, 86–116.

Wood, S. (2017 [2006]). *Generalized additive models: An introduction with R*. Boca Raton: CRC—Chapman & Hall.

Human Tongue Finite Element Model Validation with 3D MRI of subject specific phonemes articulations

Maxime CALKA[1], Pascal PERRIER[2], Yohan PAYAN[3]

[1] Sorbonne Université, Institut des Sciences du Calcul et des Données, Paris, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
[3] Univ. Grenoble Alpes, CNRS, Grenoble INP, TIMC, 38000 Grenoble, France

maxime.calka@univ-grenoble-alpes.fr, pascal.perrier@gipsa-lab.grenoble-inp.fr, yohan.payan@univ-grenoble-alpes.fr

Introduction. Biomechanical models of the human tongue integrated in the vocal tract have been developed by many research groups (Buchaillard et al., 2009; Fang et al., 2009; Stavness et al., 2011) to study speech production and speech motor control, with the simulation of tongue movements through muscle activations. The objective of this paper is to address the question of model accuracy by questioning the capacity of the model to realistically simulate speaker-specific articulation. By improving our previous finite element (FE) model of the tongue (Hermant et al., 2017), in terms of morphology, FE representations, anatomical implementation and modeling of muscles, we investigated muscle activation which enable the model to satisfactorily reproduce actual articulations of speech sounds by the reference human subject used to build the model. These articulations were measured through magnetic resonance (MR) 3D images of phonemes.

Methods. As compared to our previous model (Buchaillard et al., 2009; Hermant et al., 2017) the description of the tongue morphology has been improved with the creation of the sub-apical region, the repositioning of the external branches of the styloglossus on the styloid process and the enlargement of the posterior triangular branches of the hyoglossus muscle, which now insert on the greater horns of the hyoid bone. A mesh convergence study has also been conducted, thus providing a mesh made of 41600 tetrahedral elements with 61117 nodes. Muscle force generation has been modeled with an active transverse isotropic law based on the work of Nazari et al. (2011, 2022a) allowing active stress within an element along two different directions simultaneously. The MRI data of the articulation of vowel /i/ have been used to improve our functional partitioning of the genioglossus. The Yeoh constitutive law experimentally determined by Gerard et al. (2005) has been used to model passive tissues. The phonemes MRI have been obtained in a steady state manner. In this study the vowels /i/, /u/ and consonant /t/ are used to evaluate the model. For each of these sounds, model activations have been determined in a first step for the main muscles (different parts of the genioglossus, styloglossus, hyoglossus) starting from Buchaillard et al.'s (2009) suggestions. In a second step the activations of all the muscles were adapted step by step in order to get a reasonable approximation of the tongue shapes.

Results. In this section we only focus on the final simulations of speech articulations and the comparisons with experimental data. Figure 1 superimposes the contours of the tongue and oral cavities simulated with our model with the corresponding MR mid-sagittal, coronal and axial views. The styloglossus, the superior longitudinalis and the posterior genioglossus muscles were recruited to produce the vowel /u/, while the posterior and anterior parts of the genioglossus and the transversalis muscle were activated to produce vowel /i/. As concerns consonant /t/, the three parts of the genioglossus muscle (anterior, medium and posterior) are activated in conjunction with the superior longitudinalis so that contacts in the alveolar region can be obtained.

Discussion. As can be seen on the figure, our model is capable of generating complex shapes of the tongue in 3D space, with discrepancies to the MRI data that remain small. This is, to the best of our knowledge, the first time that tongue shapes generated with a finite element model are quantitatively compared with 3D MR data.
Some improvements will however have to be provided in the posterior part of the tongue, where the posterior genioglossus muscle is sometime not able to sufficiently compress the tongue (see Figure 1 for vowels /u/ and /i/). Dividing this muscle in two other parts might be a solution to this limitation. The constitutive law chosen for passive tissue will also have to be discussed, in particular with regard to the recent experimental uni-axial tensile tests provided by Nazari and colleagues (2022b) on human tongue tissues.
Finally, this new version of our model needs to be evaluated on other French phonemes for which MR images have been collected. The model will also be used to simulate tongue movements that will be compared to mid-sagittal trajectories already collected with electromagnetic articulography (EMA) on our reference subject.

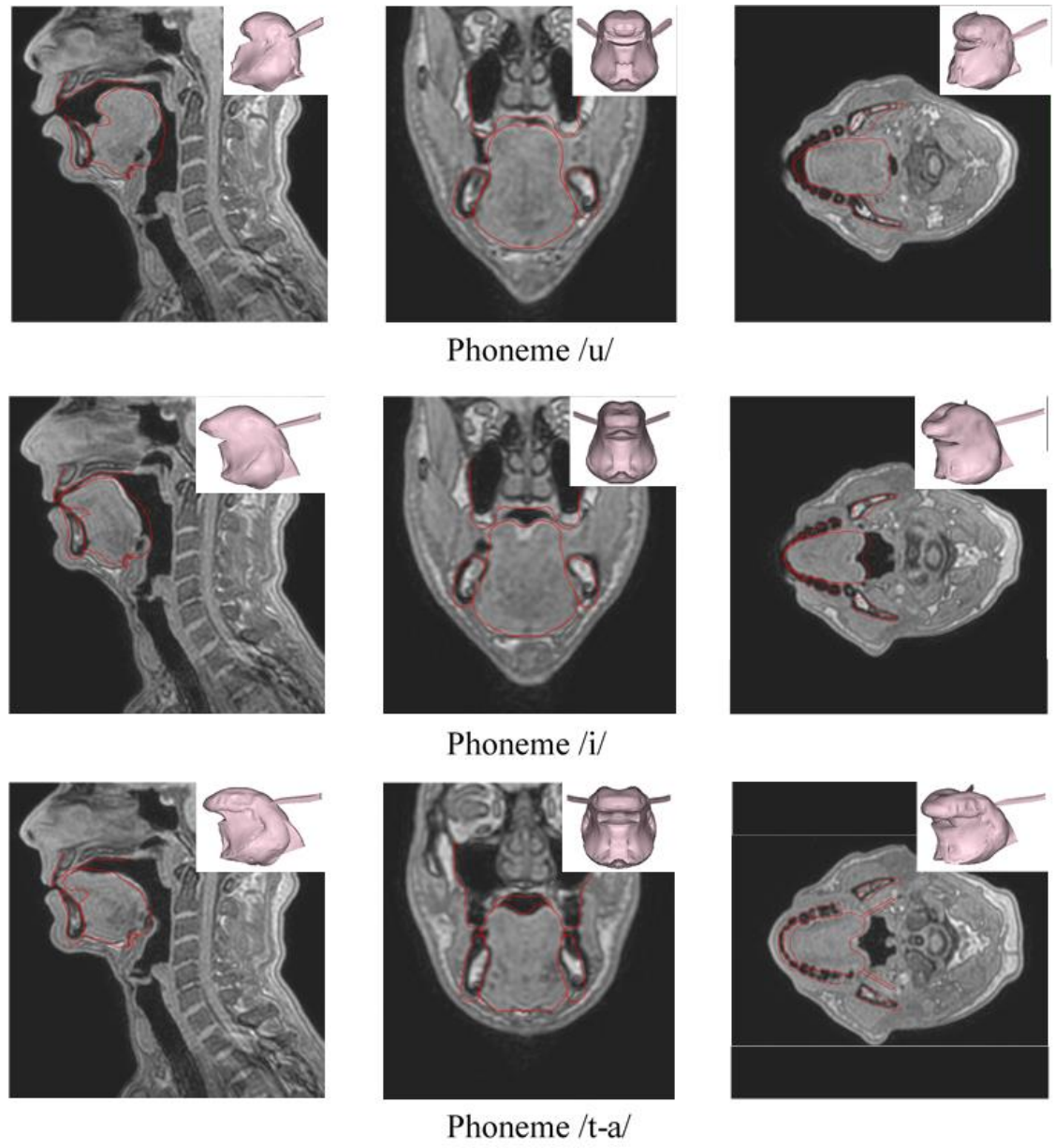


Figure 1: *Phonemes /u/, /i/ and /t/ (in context /t-a/). Contours of the tongue model (in red) in its final position superimposed with mid-sagittal, coronal and axial MR slices. 3D views of the tongue are added in sub-panels.*

References

Buchaillard, S., Perrier, P., & Payan, Y. (2009). A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning. *The Journal of the Acoustical Society of America*, *126*(4), 2033-2051.

Fang, Q., Fujita, S., Lu, X., & Dang, J. (2009). A model-based investigation of activations of the tongue muscles in vowel production. *Acoustical Science and Technology*, *30*(4), 277-287.

Gérard, J. M., Ohayon, J., Luboz, V., Perrier, P., & Payan, Y. (2005). Non-linear elastic properties of the lingual and facial tissues assessed by indentation technique: Application to the biomechanics of speech production. *Medical engineering & physics*, *27*(10), 884-892.

Hermant, N., Perrier, P., & Payan, Y. (2017). Human tongue biomechanical modeling. In *Biomechanics of Living Organs* (pp. 395-411). Academic Press.

Nazari, M. A., Perrier, P., Chabanas, M., & Payan, Y. (2011, July). A 3D finite element muscle model and its application in driving speech articulators. In *ISB 2011-XXIII Congress of the International Society of Biomechanics 2011* (pp. Paper-ID).

Nazari, M. A., Perrier, P., & Payan, Y. (2022a). Interwoven muscle fibers: a 3D two-fiber muscle active model. In *47th congress of the Society of Biomechanics, Computer Methods in Biomechanics and Biomedical Engineering* (Vol. 25, No. sup1, pp. S226-S228).

Nazari, M. A., Perrier, P., Jeannin, C., Veyre, S., Masri, C., & Payan, Y. (2022b). Ex-vivo human tongue muscle mechanical characterization. In *27th Congress of the European Society of Biomechanics (ESB'2022)*.

Stavness, I., Lloyd, J. E., Payan, Y., & Fels, S. (2011). Coupled hard–soft tissue simulation with contact and constraints applied to jaw–tongue–hyoid dynamics. *International Journal for Numerical Methods in Biomedical Engineering*, *27*(3), 367-390.

Variability in Czech children's sibilant fricative production

Tanja Kocjančič[1,2], Kateřina Vitásková[3], Kateřina Bujoková[1], Tomáš Bořil[1]

[1]Faculty of Arts, Charles University, Prague, Czech Republic
[2]Faculty of Education, University of Ljubljana, Slovenia
[3]Faculty of Education, Palacký University Olomouc, Czech Republic
tanja.kocjancicantolik@ff.cuni.cz, katerina.vitaskova@upol.cz,
katnira.b@gmail.com, tomas.boril@ff.cuni.cz

Introduction. The current pilot study aimed to explore the variability in Czech children's sibilant fricative /s, z, ʃ, ʒ/ production which is typically reported as one of the main reasons for children entering speech-language therapy. During a typical assessment protocol, the child's productions are evaluated perceptually by a speech-language therapist and compared to a standard adult variant. If the child's production does not match the standard and the age of acquisition for the specific speech sound was already achieved, the child is diagnosed with a speech sound disorder. According to a large cross-linguistic review, 75 – 85 % of children acquire /s, z, ʃ, ʒ/ at age 3;0 – 3;11, while 90 – 100 % of children acquire /s, z, ʃ/ by age 4;0 – 4;11 and /ʒ/ by age 5;0 – 5;11 (McLoed & Crowe 2018). If distortions are observed, they are typically classified based on articulatory deviations: interdental, addental, and lateral (Neubauer 2018). However, this traditional approach represents several challenges. Perceptive assessment is known to be affected by different listener-dependent factors such as the knowledge about the articulations, and auditory processing errors, such as misshearings (Kent 1996). No articulatory information about Czech sibilant fricative in children is currently available. Consequently, the currently used system of distortion classification does not cover all the possible productions and is not based on detailed articulatory analysis. Finally, children's fricative production differs from adults' and is more varied (Mass & Mailend 2017; Munson 2004), and it may be more appropriate to use children's variants as the assessment criteria (Cleland *et al.* 2018). A better understanding of how children produce sibilant fricatives could improve the assessment by allowing a speech-language therapist to make more informative conclusions about the underlying articulation based on the perceptive analysis. This would lead to a selection of more appropriate therapy goals. If the therapist correctly identifies the problematic element of the articulation (e.g., the tongue has a correct shape but is placed further back in the oral cavity), only this element can be addressed in the therapy and the articulatory instructions given to the child can be significantly more focused. Problems with understanding and executing even simple articulatory movements have been shown for adults (Ouni 2014) and it can be expected that children would perform at least similarly, if not worse, due to a developing motor control system. To address this clinical application, we must first better understand the productions of typically developing children.

Methods. 11 children aged 3 to 5 years participated in the study. Due to poor visibility of the tongue contour in the ultrasound recording, only six children are presented here: CH3-1 (F, 3;5), CH3-2 (M, 3;10), CH4-1 (F, 4;5), CH4-2 (M, 4;10), CH5-1 (F, 5;1), CH5-2 (M, 5;11). All were monolingual Czech speakers attending the same kindergarten and had no known impairments affecting speech. The children made four repetitions of a word list consisting of six disyllable words per target sibilant /s, z, ʃ, ʒ/ (6 words x 4 sibilants x 4 repetitions = 96 production). The targets were in a CV (V = /i, a, u/) and C1C2V (C2 = /p, t, k/ or /b, d, g/; voicing matched with C1) context. Words were elicited via an imitative picture-naming task (Edwards & Beckman 2008) which made sure that all children reliably produced the targets. Audio and ultrasound recordings of midsagittal tongue contour were made with the Micro system (Articulate Instruments Ltd., 20012) and the probe stabilization headset (Articulate Instruments Ltd. 2008) at the kindergarten. The obtained data was first segmented, and the targets were transcribed based on the perceptive analysis and visual inspection of the waveform and spectrogram. To account for the variability, the errors were further marked in terms of change in voicing or place of articulation, presence of lateral airflow or weak articulation, and their combinations. The tongue surface was then traced in the middle of the target segment duration. For each child and each of the four targets, a subset of data where the production matched the target was selected, and mean tongue contour and standard deviations were calculated for these subsets. Such representation of absolute tongue position and shape allowed observing lingual stability in repetition.

Results. Perceptive analysis revealed high variability between and within individual children. Ordered from the youngest to the oldest participant, the children produced correctly 96%, 13%, 83%, 75% 0% and 71% of /s/ targets, 13%, 22%, 65%, 87%, 13% and 42% of /z/ targets, 83%, 0%, 71%, 22%, 50% and 58% of /ʃ/ targets and 8%, 0%, 38%, 13%, 25% and 21% of /ʒ/ targets. All except CH5-1 showed a higher % of correctly produced alveolar than palatal targets (in pairs matched by voicing), with /ʒ/ being the least correct, and 4/6 for voiceless than voiced sounds (across both places of articulation). Children in all age groups produced more error types for voiced (3 – 12 different error types per child) than for voiceless (1 – 4 error types per child) sibilants, with /ʒ/ showing the most types for 5/6 children. The most error types were observed in the speech of the two 5-year-olds: CH5-2 showed 9 error types for /z/ and 12 for /ʒ/. The most frequent errors resulted from changes in voicing and horizontal tongue placement. Visual inspection of ultrasound data of

production matching the targets (3 – 22 per sibilant and participant) showed greater lingual stability, particularly in the front part of the tongue, for the youngest speakers, with a decrease with age, as well as greater stability for palatal than for alveolar place of articulation. Figure 1 shows the mean and standard deviation based on the subset of correctly produced targets for children CH3-1, CH4-1, and CH5-2. The youngest child CH3-1 seemed to make narrower (along the length of the tongue) linguopalatal contact for /ʃ, ʒ/ than the older two.

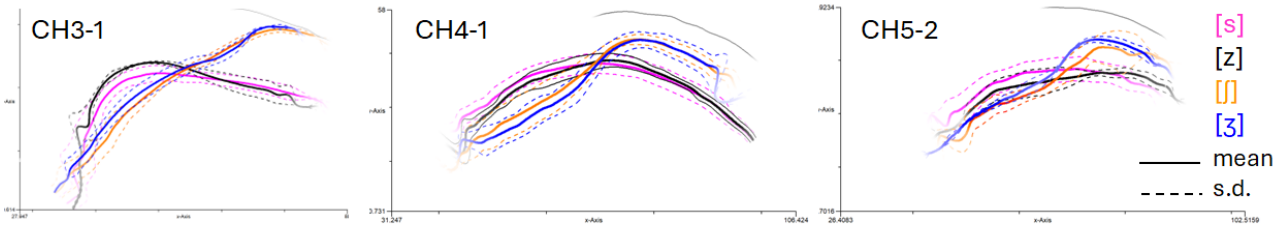


Figure 1: *Mean and standard deviation midsagittal tongue contours for /s, z, ʃ, ʒ/ for CH3-1 (left) CH4-1 (middle) and CH5-2 (right). The tongue front is on the right side of the plots, palate is in gray.*

Discussion. Overall, the results show a high within- and between-speaker variability in the attested sibilant fricatives in all explored measures: % of correct productions, number of error types, and tongue placement in the perceptually correct productions. No apparent decrease with age was noted. This is in accordance with Zharkova (2018) who showed that lingual variability decreases after the age of 5 years. In terms of age of acquisition, only the 5-year-olds are expected to master the production of /s, z, ʃ/, however, they did not differ from the youngest children. The normative data most commonly results from a single production of 1 – 3 words, whereas in the current study, the children produced four repetitions of 6 words per target sound. Considering known variability in child speech, it may be more appropriate to base norms on larger and repeated word lists. The results on % of correct productions and number of error types revealed that children have more problems producing voiced sibilant fricatives, particularly /ʒ/. An earlier EPG study on the same sounds in Czech adults has revealed a larger tongue-palate contact in voiced than in voiceless targets (Skarnitzl *et al.* 2013). The authors discussed that lower articulatory precision is needed for voiceless sounds. It can be expected that in children a lower precision demand, coupled with an immature motor control system, results in a greater likelihood of erroneous productions, as shown in the data. Children were more successful in producing alveolar than palatal fricatives (matched by voicing), most likely due to the former having a clear contact point: the tip of the tongue in contact with the lower incisor. Current results also show that older children employed a larger part of the front of the tongue in /ʃ, ʒ/, suggesting that this observation could be used in evaluating the maturation of lingual patterns in sibilant fricative production. Finally, the current small data set will be increased over the following months and articulatory data will be analyzed quantitatively.

This work was supported by the European Regional Development Fund project " Beyond Security: Role of Conflict in Resilience-Building" (reg. no.: CZ.02.01.01/00/22_008/0004595) and by the Czech Science Foundation Grant No. 23-05494S.

References

Articulate Instruments Ltd. (2008). Ultrasound stabilisation headset users manual: Revision 1.4. Articulate Instruments Ltd.

Articulate Instruments Ltd. (2012). Articulate Assistant Advanced user guide: Version 2.14. Articulate Instruments Ltd.

Cleland, J., Wrench, A., Lloyd, S., & Sugden, E. (2018). ULTRAX2020: Ultrasound technology for optimising the treatment of speech disorders: Clinicians' Resource Manual.

Edwards, J., & Beckman, M. E. (2008). Methodological questions in studying consonant acquisition. *Clinical linguistics & phonetics*, 22(12), 937-956.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders.

American Journal of Speech-Language Pathology, 5(3), 7-23.

Maas, E., & Mailend, M. L. (2017). Fricative contrast and coarticulation in children with and without speech sound disorders. *American journal of speech-language pathology*, 26(2S), 649-663.

McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American journal of speech-language pathology*, 27(4), 1546-1571.

Munson, B. (2004). Variability in/s/production in children and adults. *Journal of speech, language and hearing research*, 47, 58-69.

Neubauer, K. (2018). Vývojové a přetrvávající poruchy artikulace a fonologického rozlišování hlásek. In: K.Neubauer (Ed), Kompedium klinické logopedie: diagnostika a terapie poruch komunikace. Praha: Portal.316-341.

Ouni, S. (2014). Tongue control and its implication in pronunciation training. *Computer Assisted Language Learning*, 27(5), 439-453.

Skarnitzl, R., Šturm, P. & Machač, P. (2013). The phonological voicing contrast in Czech: An EPG study of phonated and whispered fricatives. In: Proceedings of Interspeech 2013, 3191–3195.

Zharkova, N. (2018). An ultrasound study of the development of lingual coarticulation during childhood. *Phonetica*, 75(3), 245-271.

The role of the supplementary motor area in speech production: Evidence from participants who do, and do not stutter.

Charlotte E. E. Wiltshire,[1,2], Nicole Benker[2], Rosa Hufschmidt[2], Anton Gadringer[2], Philip Hoole[2]

[1]Department of Psychology, Bangor University, Wales, UK
[2]Institute of Phonetics and Speech Processing, Ludwig-Maximilians-University, Munich, Germany.
c.wiltshire@bangor.ac.uk, nicole.benker@campus.lmu.de,
rosa.hufschmidt@campus.lmu.de, hoole@phonetik.uni-muenchen.de

Introduction. The basal-ganglia-thalamo-cortical network is thought to underlie the coordination of speech and non-speech movements. Neuroimaging studies have revealed anatomical and functional differences in this network in people who stutter (Alm 2004; Frankford et al., 2021; Chang & Guenther, 2020). Stuttering is markedly reduced when speaking with an external cue (such as a metronome) compared with internally cued speech (e.g. conversational speech). Two neural loops may explain this "rhythm effect": An "internal timing network" comprising a basal-ganglia-SMA loop and an "external timing network" comprising a pre-motor-basal-ganglia-cerebellum loop (Alm, 2004). In this study, we examined the hypotheses that for both people who do, and do not, stutter, the SMA is 1) involved in the coordination of speech movements and 2) is particularly sensitive to internally cued speech. We further hypothesise that these effects will be strongest in people who stutter, representing differences in the underlying function of the basal-ganglia-SMA motor loop. The rationale and methodology have been pre-registered as part of an accepted Stage 1 registered report at *Brain Communications* (available: https://osf.io/hpve5/).

Methods. Twenty-one participants who stutter and 21 who do not stutter completed two sessions. For each session, repetitive Transcranial Magnetic Stimulation (rTMS; 0.6 Hz, 15 minutes) was used to disrupt the function of the SMA or the hand representation of the primary motor cortex (Hand-M1; control site) for a further 15 minutes after stimulation. Before and after rTMS, Electromagnetic Articulography was used to record speech movements whilst participants repeated simple speech sequences (e.g., "bi da gu"). We aimed to create conditions as close to the opposing ends of the internal and external cueing spectrum as possible whilst retaining the experimental control needed for EMA studies. In the external condition, participants repeated the speech sequences with a metronome whilst the text appeared on a screen. In the internal condition, participants viewed the sequences, then produced them without a metronome and with a blank screen. Speech motor variability was calculated using the coefficient of variation of the area under the curve of each utterance. This approach captures variability in both the amplitude and duration of speech movements. Motor Evoked Potentials were elicited using single pulse monophasic TMS over Hand-M1 before and immediately after the repetitive TMS to measure changes in cortical excitability.

Results. There was no difference in variability between the two groups at baseline (i.e. before stimulation). In addition, there was no difference in baseline performance between the first and second sessions, showing good test-retest reliability. rTMS applied to the Hand-M1 successfully reduced the amplitude of motor evoked potentials elicited from the hand muscle (p=.001). rTMS applied to the SMA did not reduce the amplitude of motor evoked potentials in the hand muscle, as expected. We assume, therefore, that the stimulation protocol was successful at reducing cortical excitability in both the Hand-M1 condition (control site) and SMA and that these effects are focal to the brain area targeted. There was no change in kinematic variability for both internal and external speech conditions following repetitive TMS to the SMA or Hand-M1, see Figure 1.

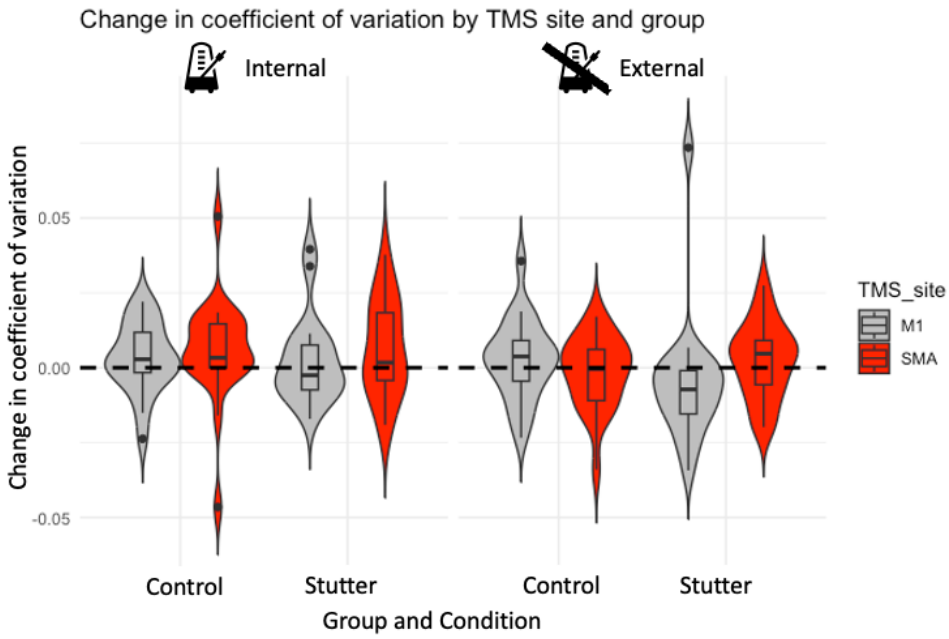


Figure 1: Change in variability (coefficient of variation) from pre- to post-stimulation. M1 = hand representation of the primary motor area, SMA = Supplementary Motor Area. Horizontal dotted line at zero represents no change.

Discussion. Firstly, these results do not reproduce previous findings that people who stutter have greater speech motor variability during fluent speech production compared with a control group (Wiltshire et al., 2021; 2023). This may be explained by a difference in kinematic recording techniques: Electromagnetic articulography, as used here, has higher spatial and temporal resolution compared with vocal tract MRI but requires small electrodes be attached to the surface of the articulators, thus altering typical sensory-motor feedback. This disruption is likely to be particularly important for people who stutter and may reduce our ability to detect subtle between-group differences. Secondly, despite successfully targeting the cortex using rTMS, stimulating the SMA did not impair speech motor control during internally or externally cued speech production in both people who stutter and the control group. It may be that we were wrong in our prediction that the SMA is particularly important for internally generated speech compared with externally cued speech. Another explanation is that even though the SMA-Basal Ganglia motor loop was disrupted, the alternate cerebellum-basal ganglia motor loop was sufficient to maintain low levels of variability in both the internal and external conditions. Additional, more detailed analyses are ongoing that leverage the excellent spatial and temporal resolution of electromagnetic articulography. These analyses aim to determine the impact of initiating a sequence (1st syllable of the sequence) and the effect of duration and amplitude of speech production.

References

Alm, P. A. (2004). Stuttering and the basal ganglia circuits: A critical review of possible relations. *Journal of Communication Disorders*, *37*(4), 325–369

Chang, S.-E., & Guenther, F. H. (2020). Involvement of the Cortico-Basal Ganglia-Thalamocortical Loop in Developmental Stuttering. *Frontiers in Psychology*, (10)

Frankford, S. A., Heller Murray, E. S., Masapollo, M., Cai, S., Tourville, J. A., Nieto-Castañón, A., & Guenther, F. H. (2021). The neural circuitry underlying the "rhythm effect" in stuttering. *Journal of Speech, Language, and Hearing Research*, *64*(6S), 2325-2346

Wiltshire, C. E. E., Chiew, M., Chesters, J., Healy, M. P., & Watkins, K. E. (2021). Speech Movement Variability in People Who Stutter: A Vocal Tract Magnetic Resonance Imaging Study. *Journal of Speech, Language, and Hearing Research*. *64(7)*, 2438-2452

Wiltshire, C. E. E., Cler, G. J., Chiew, M., Freudenberger, J., Chesters, J., Healy, M. P., Hoole, P., & Watkins, K. E. (2023). Speaking to a metronome reduces kinematic variability in typical speakers and people who stutter. *Preprint, OSF* https://osf.io/fj3un

Task effects and phonological error patterns in Australian English-Dutch bilingual children

Hayo Terband[1], Bhavana Bhat[1], Anniek van Doornik[2,3]

[1]Department of Communication Sciences and Disorders, University of Iowa, Iowa City IA, USA
[2]HU University of applied Sciences, Utrecht, The Netherlands
[3]Institute for Language Sciences, Utrecht University, Utrecht, The Netherlands
hayo-terband@uiowa.edu, bhavana-bhat.uiowa.edu, anniek.vandoornik@hu.nl

Introduction. Determining whether suspected speech 'abnormalities' in bilingual children are due to a pattern of bilingual language acquisition - and language dominance - or due to a speech sound disorder is a challenging task for speech-language pathologists (SLPs; Girolametto & Cleave 2010; McLeod, et al. 2013). With the advantage of not directly involving linguistic knowledge, nonword imitation (NWI) is often used as a diagnostic task, in particular with respect to phonological and language disorders (e.g., Boerma et al. 2015; Chiat & Roy 2007; Dos Santos & Ferré 2018; Kapalková et al. 2013; Ortiz 2021; Schwob et al. 2021). However, previous studies have shown that 5-12 year old typically developing bilingual children can score significantly worse on these tasks compared to typically developing monolingual children (Engel de Abreu 2011; Thordardottir & Brandeker 2013; Windsor et al. 2010). Furthermore, studies with bilingual children have reported only general outcome measures such as percentage of items or phonemes correct (e.g., Ortiz 2021; Schwob et al. 2021). The aim of the present study was to investigate how NWI productions of English-Dutch bilingual children differ from other speech tasks and from norm data, as to establish the potential role of NWI in diagnosing speech sound disorders in bilingual children.

Methods. 77 typically developing Australian English-Dutch bilingual children ranging between 4 and 12 years of age (M=7.96, SD=2.40; 43 girls, 34 boys) participated in this study. All children attended a regular Australian school during the week and additionally attended the Dutch language school in Sydney for appr. 2 hours each weekend, for which they had about 2 hours of homework.
All children completed the Dutch test battery *Computer Articulation Instrument* (CAI; Maassen et al. 2019), which includes picture naming (PN), nonword imitation (NWI), consistency of 5 consecutive repetitions of words and nonwords (WR & NWR) and diadochokinesis maximum repetition rate (MMR). Data on language exposure was collected through parent/caregiver questionnaires. 65% speak more than half of the time Dutch at home, 20% speak more than half of the time a combination of English and Dutch at home, 9% speak more than half of the time English at home, 6% speak more than half of the time a combination of Dutch and another language at home.
The CAI is normed for children up to 7 years old, but in this case also used with older children. For the purpose of the analysis, the children were split in two age groups: 4-7 years old (<7 years; n=29) and 7-12 years old (≥7 years; n=48).
Group-level quantitative phonological error analyses compared performance across tasks while qualitative error analyses investigated the error patterns in terms of phonological processes. Error patterns were analyzed acoustically to inform their interpretation. Possible associations between tasks were examined through a correlational analysis, both with the raw data (across age-groups) and with z-scores (per age-group). Correlations were calculated by means of Pearson's r.

Results. The English-Dutch bilingual children scored lower compared to the norm data for percentages consonants (PCCI) and vowels correct (PVC) on PN and consistency (WR & NWR) while PCC and PVC on NWI, and MMR were age appropriate (**Figure 1**). The correlation analysis revealed a positive correlation between scores on NWI and NWR consistency. No other significant correlations were observed.
In PN, the phonological processes fronting, devoicing and gliding occurred the most, however, only devoicing occurred more often compared to the norm. The most common phonological processes in NWI were fronting, dentalization, voicing, devoicing and gliding, but these did not occur more compared to the norm. Stopping of fricatives and h-zation did occur more compared to norms in NWI.

Discussion. These results confirm NWI as (most) language-neutral assessment of speech production in bilingual children. In terms of phonological processes, the results suggest interference of English phonology and a loss of readily available phonological representations including motor goals for Dutch speech sounds. In PN for example, the children produced relatively many syllable-initial voiced plosives with an English VOT (voice onset time), which in Dutch maps onto the voiceless cognate. Interestingly, excessive devoicing did not occur in NWI, meaning that the children were able to perceive segments as pre-voiced and produce them accordingly. Similarly for fricatives, a pattern was found in which productions in PN with English COG (spectral center of gravity) map onto a Dutch phoneme, while (failed) attempts to

match the specific COG in NWI resulted in stopping or h-zation. Acoustic measurements of the produced VOT's and COG's confirm this explanation.
The correlation between NWI and consistency NWR indicates that the children who produced less errors were also more consistent. Upon closer inspection, additionally, the results on the consistency tasks showed an interesting pattern of increased transfer of English features with each subsequent repetition (e.g., *telefoon*: /teləfoːn/ > [tɛləfoːn] > [tɛləfɔn] > [tæləfɔn] > [t^hæləfɔn]). The memory trace of the acoustic model appears to fade with each repetition and the task thus slowly becomes a delayed imitation task. In conclusion, VOT, fricatives, and vowels need attention of SLPs assessing English-Dutch bilingual children.

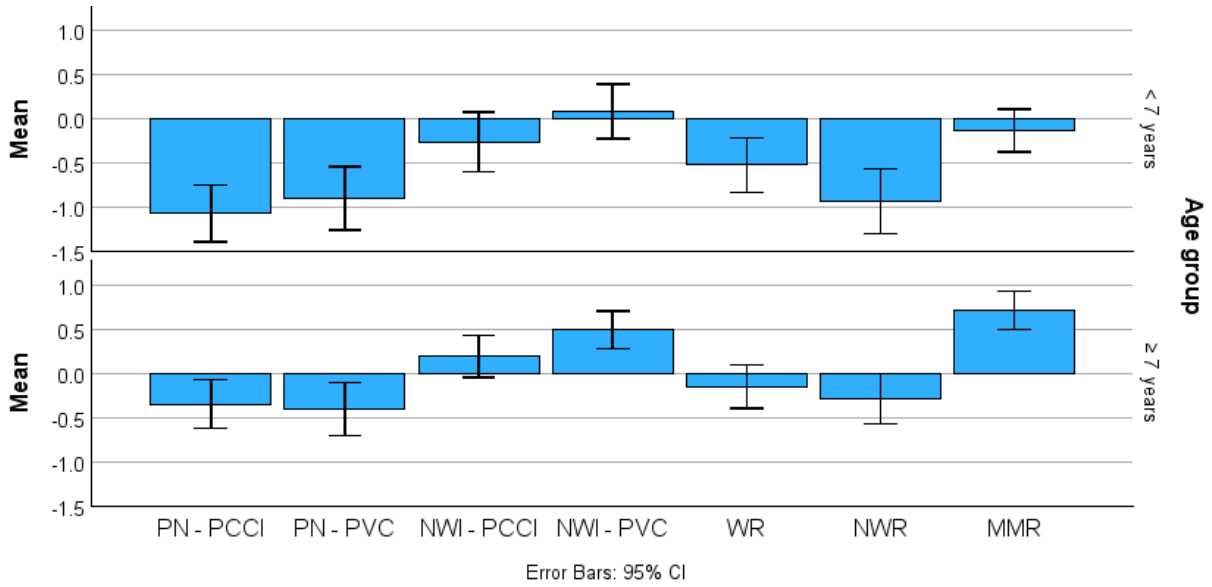


Figure 1: *Mean z-scores on the picture naming (PN), nonword imitation (NWI), consistency (WR = word repetition; NWR = nonword repetition), and diadochokinesis (MMR = maximum repetition rate) tasks, broken down by age group (PCCI = percentage consonants correct in syllable-initial position; PVC = percentage vowels correct).*

References

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). "A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language". In: *Journal of Speech, Language, and Hearing Research*, 58.6, pp. 1747-1760.

Chiat, S., & Roy, P. (2007). "The preschool repetition test: An evaluation of performance in typically developing and clinically referred children". In: *Journal of Speech, Language, and Hearing Research*, 50.2, pp. 429-443.

Dos Santos, C., & Ferré, S. (2018). "A nonword repetition task to assess bilingual children's phonology". In: *Language Acquisition*, 25.1, pp. 58-71.

Engel de Abreu, P. M. (2011). "Working memory in multilingual children: Is there a bilingual effect?" In: *Memory*, 19.5, pp. 529-537.

Girolametto, L., & Cleave, P. L. (2010). "Assessment and intervention of bilingual children with language impairment". In: *Journal of communication disorders*, 43.6, pp. 453-455.

Kapalková, S., Polišenská, K., & Vicenová, Z. (2013). "Non-word repetition performance in Slovak-speaking children with and without SLI: novel scoring methods". In: *International Journal of Language & Communication Disorders*, 48.1, pp. 78-89.

Maassen, B., van Haaften, L., Diepeveen, S., van den Engel-Hoek, L., Veenker, T., Terband, H., & De Swart, B. (2019). "Computer Articulatie-Instrument (CAI)." Amsterdam: Boom test uitgevers.

McLeod, S., Verdon, S., Baker, E., Ball, M. J., Ballard, E., David, A. B., ... & Zharkova, N. (2017). "Tutorial: Speech assessment for multilingual children who do not speak the same language (s) as the speech-language pathologist". In: *American Journal of Speech-Language Pathology*, 26.3, pp. 691-708.

Ortiz, J. A. (2021). "Using nonword repetition to identify language impairment in bilingual children: A meta-analysis of diagnostic accuracy". In: *American Journal of Speech-Language Pathology*, 30.5, pp. 2275-2295.

Schwob, S., Eddé, L., Jacquin, L., Leboulanger, M., Picard, M., Oliveira, P. R., & Skoruppa, K. (2021). "Using nonword repetition to identify developmental language disorder in monolingual and bilingual children: A systematic review and meta-analysis". In: *Journal of Speech, Language, and Hearing Research*, 64.9, pp. 3578-3593.

Thordardottir, E., & Brandeker, M. (2013). "The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores". In: *Journal of Communication Disorders*, 46.1, pp. 1-16.

Windsor, J., Kohnert, K., Lobitz, K. F., & Pham, G. T. (2010). "Cross-Language Nonword Repetition by Bilingual and Monolingual Children". In: *American Journal of Speech-Language Pathology*, 19.4, pp. 298-310.

Compensatory Strategies in Individuals with Moebius Syndrome: A Case Study

Anne Hermes[1], Ivana Didirková[2], Philipp Buech[1], Gilles Vannuscorps[3]

[1]Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne Nouvelle, Paris, France
[2]UR 1569 TransCrit, Université Paris 8 Vincennes – Saint-Denis, France
[3]Psychological Sciences Research Institute/IONS, Université catholique de Louvain, Belgium
anne.hermes@sorbonne-nouvelle.fr, ivana.didirkova@univ-paris8.fr,
philipp.buech@sorbonne-nouvelle.fr, gilles.vannuscorps@uclouvain.be

Introduction. Moebius syndrome is a rare (0.00002 to 0.002% of the world's population) congenital neuromuscular disorder characterized by the absence or underdevelopment of the 6th and 7th cranial nerves, which control horizontal eye movements and facial muscles (Verzijl et al. 2005). Key articulatory features in individuals with Moebius syndrome thus include lip paralysis and incomplete lip closure. These characteristics can affect, among others, the production of bilabial (/m, b, p/) and labio-dental (/f, v/) consonants, and rounded vowels, which all recruit the upper and lower lips and involve the jaw. Previous studies have reported that a large proportion of these individuals present with frequent misarticulations related to impaired labial function (Kahane 1979, Helmick 1980, Murdoch et al. 1997, Sjögreen et al. 2001, del Carmen Pamplona et al. 2022). Interestingly, however, about 20% of these individuals appear to succeed in developing particularly efficient compensatory articulatory movement (Sjögreen et al. 2022). A detailed description of these compensatory articulatory patterns has the potential to inform fundamental issues in articulatory phonetics, specifically speech motor control, and to guide speech therapists in their attempts to improve the intelligibility of patients with congenital and acquired lip paralysis. However, to date, research focusing specifically on a quantitative analysis of speech motor control strategies in individuals with Moebius syndrome with intact intelligibility does not exist. This study aims to fill this gap by providing the first detailed acoustic and articulatory analysis of speech motor control mechanisms in two individuals with complete lip paralysis but intact speech intelligibility.

Methods. We collected and analyzed acoustic and articulatory data (EMA, AG 501) from two female individuals with Moebius syndrome (S1=41 years; S2=43 years) who present with congenital bilateral facial palsy (see Fig. 1), but intact speech intelligibility as determined by a speech therapist.

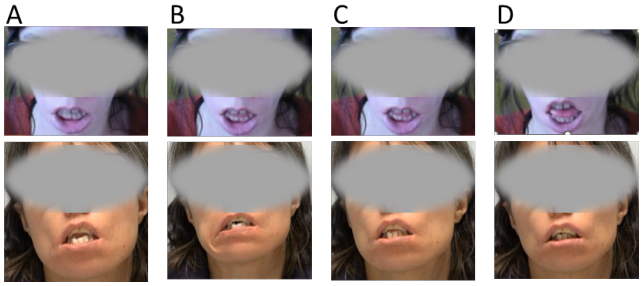


Figure 1: *S1 (top) and S2 (bottom) attempting to (A) close the lips as much as possible, (B) stretch the lips, (C) round the lips, and (D) open the lips as much as possible.*

For comparison, we also collected data from a French control speaker (female, 28 years old). Our methodology entailed the collection of syllable repetitions (DDK), sentence production, and text reading. Here, we present the first results of one of the DDK tasks (production of syllable /pa/ on one breath cycle as fast and as precise as possible), juxtaposed with the articulatory data of a control speaker. For the collection of the articulatory data, we put sensors on the vermillion border of the upper and lower lips (ULIP, LLIP), tongue tip (TTIP), tongue body (TBO), and chin (CHIN) as well as reference sensors behind the ears. The articulatory signal was sampled at a rate of 1250 Hz and filtered using a Butterworth low pass with a cut-off frequency of 25 Hz and an order of 5. The articulatory data was head-corrected and rotated to the occlusal plane. The articulatory analysis focused on the movement and coordination patterns of the tongue, lips, and jaw during the production of fast syllable repetition tasks. We applied the following measures (over ten /pa/ repetitions): (1) convex hull area in mm^2 (range/extent of a sensor movement), (2) path length in mm (total distance of a sensor movement), and (3) Euclidian distances between CHIN and LLIP sensors.

Results. Figure 2 displays the (1) convex hull area (i.e., movement extent of each sensor) during the labial DKK task (first and last /pa/ production were not taken into account). For the two individuals with Moebius syndrome (S1, S2), we can observe (i) a reduced range of motion in the movement of the lips and the jaw (Fig. 2: dashed and dotted; Table 1;

e.g., LLIP: S1=3.98mm^2; S2=4.12mm^2) compared to the control speaker C1 (LLIP=7.4mm^2), whereas the range of motions of the tongue tip and tongue body is much greater (e.g., TTIP: S1=12.46mm^2; S2=74.23mm^2) than for the control speaker (TTIP=1.97mm^2). Further, the results clearly show that the tongue tip is much higher (towards the palate) than for the control speaker.

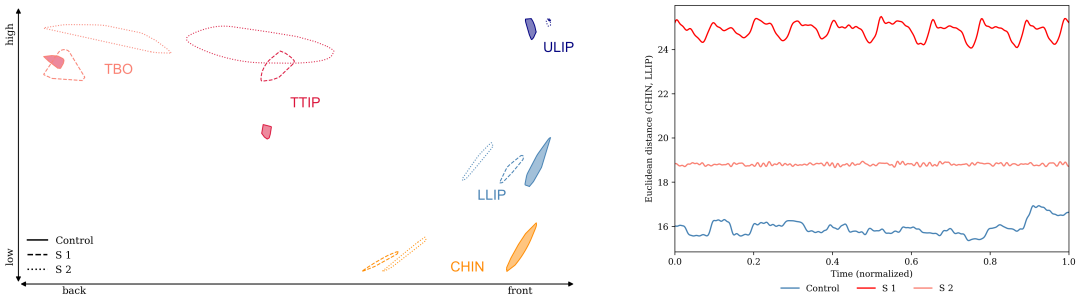


Figure 2: *(left) Convex hull area of ten /pa/ repetitions in two individuals with Moebius syndrome (S1: dashed; S2: dotted lines) and one control speaker (C1: solid lines, filled). Positions are min-max normalized for each speaker. (right) Euclidean distance between chin and lower lip sensor for C1 (blue), S1 and S2.*

As expected, the observed strategy of the individuals with Moebius syndrome is compensation for the lip paralysis by producing excessive tongue tip/body movements. The compensation is also reflected in the path length (see Table 1), where for S1 and S2 the sensors TBO and TTIP show much higher values, whereas for the control speaker higher values (longer ways) are found for ULIP, LLIP and CHIN. It shall be noted that the two individuals show also individual differences, in that e.g., S2 used a much larger area and longer path for the tongue tip movement (during the production of labial stops) than S1 did (Fig. 2, left, TTIP: convex hull area S2=74.23mm^2 vs. S1=12.46mm^2; path length S2=196.69mm vs. S1=104.5mm). This difference in strategies is further confirmed by the Euclidian distances between CHIN and LLIP sensors analysis, showing preserved movement in S1, while S2 seems to reduce her movements (Fig. 2, right). More results on coordination patterns and alternating DDK (i.e., /patakapataka, badegobadego/) will be presented at the conference.

Table 1: *Convex hull area (in mm^2) and path length (in mm) for sensor movements over ten /pa/ repetitions, comparing control (C1) with individuals with Moebius syndrome (S1, S2).*

	Convex hull area (in mm^2)					Path length (in mm)				
	TBO	TTIP	ULIP	LLIP	CHIN	TBO	TTIP	ULIP	LLIP	CHIN
C1	2.99	1.97	2.66	7.4	7.37	27.37	22.95	47.59	130.01	124.53
S1	16.4	12.46	0.07	3.98	2.29	54.19	104.5	9.42	90.34	69.88
S2	30.17	74.23	0.46	4.12	3.34	154.89	196.69	14.86	109.11	112.52

Discussion/Conclusion. The articulatory analysis of DDK in individuals with Moebius syndrome reveals that (1) the lips are not involved in the closure for the production of labial stops, whereas (2) the tongue movements indicate compensatory strategies. It shall be mentioned that DDK is an artificial movement paradigm (Ziegler et al. 2002), which, however, gives us first indications of how the speech motor control system is affected. These preliminary results on speech motor control in individuals with Moebius syndrome reveal intriguing strategies adopted by these individuals to navigate their speech limitations in producing labial sounds. This study hence provides novel, valuable insights into speech production mechanisms of individuals with Moebius syndrome which can pose a challenge to speech production models. Further, understanding these mechanisms can aid speech therapists/clinicians in developing tailored interventions and assistive technologies to enhance communication outcomes for individuals with Moebius syndrome.

References

del Carmen Pamplona, M., Ysunza, P.A., Telich-Tarriba, J., Chávez-Serna, E., Villate-Escobar, P., Sterling, M., & Cardenas-Mejia, A. (2020). Diagnosis and treatment of speech disorders in children with Moebius syndrome. *International Journal of Pediatric Otorhinolaryngology*, 138, 110316.

Kahane, J.C. (1979). Pathophysiological Effects of Möbius Syndrome on Speech and Hearing. *Arch Otolaryngol*. 105(1):29–34.

Murdoch, B.E., Johnson, S.M. & Theodoros, D.G. (1997) Physiological and perceptual features of dysarthria in Moebius syndrome: directions for treatment, *Pediatric Rehabilitation*, 1:2, 83-97, DOI: 10.3109/17518429709025851

Sjögreen, L., Andersson-Norinder, J., & Jacobsson, C. (2001). Development of speech, feeding, eating, and facial expression in Möbius sequence. *International Journal of Pediatric Otorhinolaryngology*, 60(3), 197-204.

Verzijl, H.T.F.M., Valk, J., de Vries, R. & Padberg, G.W. (2005). Radiologic evidence for absence of the facial nerve in Moebius syndrome. *Neurology, 64*, 849-855.

Ziegler, W. (2002). Task-related factors in oral motor control: speech and oral diadochokinesis in dysarthria and apraxia of speech. *Brain and Language*, *80*(3), 556–575.

Validating the Use of Simulation Based Inference for Feedback Aware Control of Tasks in Speech (FACTS) and Human F1 Compensation Data

Alvincé L. Pongos[1], Kwang S. Kim[2], Ben Parrell[3], Vikram Ramanarayanan[4,5], Jessica L. Gaines[1], Srikantan S. Nagarajan[4], John F. Houde[4]

[1]UC Berkeley – UCSF Graduate Program in Bioengineering, University of California-San Francisco, San Francisco, California, USA
[2]Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, Indiana, USA
[3]Department of Communication Sciences and Disorders, University of Wisconsin-Madison, Madison, Wisconsin, USA
[4]Department of Otolaryngology, University of California-San Francisco, San Francisco, California, USA
[5]Modality.ai, San Francisco, California, USA

alvince_pongos@berkeley.edu,kwangkim@purdue.edu, bparrell@wisc.edu, vikram.ramanarayanan@modality.ai, jessica.gaines@berkeley.edu, srikantan.nagarajan@ucsf.edu, john.houde@ucsf.edu

Introduction. Speech differences between healthy controls (HC) and disordered populations exist, but determining the underlying neural computations that explain these differences remains a challenge. For example, compared to healthy controls, those with cerebellar degeneration (CD) may show a heightened compensation response to F1 perturbations (Parrell et al., 2017, though cf Parrell et al., 2021), but the hypothesis space for what could explain these differences remains relatively unconstrained and uncertain. Simulation-based inference (SBI) offers a principled approach to estimate posterior distributions over mechanistic model parameters (Cranmer et al., 2020), thereby quantifying parameter certainty and taking a first step towards generating explanatory hypotheses describing behavioral differences between groups. However, efforts to date that apply SBI to speech data have been restricted to models of laryngeal control. (Gaines et al., in prep). Articulatory models of speech production typically involve a larger number of parameters that may interact in complex ways. For example, the FACTS model (Parrell et al., 2021; Kim et al., 2023) has at least 9 tunable parameters that may interact strongly, and it is possible that the search space may be difficult for SBI to converge on an accurate solution for all parameters simultaneously. Here, we validate the SBI approach by 1) recovering known model parameters from a simulated dataset and 1) estimating parameters that qualitatively reproduce human behavior in response to an external perturbation of perceived vowel formants.

Methods. FACTS employs a hierarchical state feedback control architecture to control simulated vocal tract gestures and production of intelligible speech (Parrell et al. 2021). Among many speech phenomena, FACTS is able to model the behavioral response to external perturbations applied to vowel formants. Here, we tested the ability of simulation-based inference (SBI), a likelihood-free inference method that estimates posterior distributions of model parameters, to recover the FACTS parameters that produce these compensatory responses, both in a FACTS-simulated dataset and in human behavioral data from neurobiologically healthy speakers (Parrell et a. 2017). SBI is an alternative method to standard Bayesian approaches because it does not rely on a likelihood function (Cranmer et al., 2020). SBI is useful when likelihoods are intractable to estimate and scientists have access to a simulator of observed phenomena. That is SBI estimates a posterior distribution $p(\theta|\boldsymbol{x})$ over a set of parameters θ given observations $\boldsymbol{x}$ and a simulator (i.e. a computational model). The Automatic Posterior Transformation (APT) SBI method is used in this work over alternatives because it has been shown to be more flexible, scalable and efficient than previous simulation-based inference techniques (Greenberg et al. 2019).

We carried out two assessments. In the first assessment, we examined whether SBI could recover model parameters based on FACTS simulation (i.e,. SBI-estimated F1 trajectories vs. original FACTS F1 trajectories). For SBI-estimated F1 trajectories, the "ground-truth" FACTS parameters were first taken from Kim et al. (2023) and uniform prior distributions were then estimated by performing parameter sweeps along each dimension until boundaries were found that contained stable, plausible speech behavior. The uniform priors were then used to generate 100,000 (input parameters, F1 trajectory) pairs using the newest version of FACTS (Kim et al. 2023). This large dataset (100,000 pairs of F1 trajectories and their corresponding control parameter values) provided a large, uniformly distributed search space that trained an estimate of $p(\theta|\boldsymbol{x})$. The ground-truth F1 trajectory of FACTS was then used $p(\theta|\boldsymbol{x}_{\text{FACTS}})$ to estimate joint and marginal posterior distributions over the parameters. The modes of the marginals were input to FACTS to generate the SBI-estimated-F1 trajectories, which were compared with the original "ground-truth" FACTS F1 trajectory generated by FACTS parameters from Kim et al. (2023)

In our second assessment, we investigated whether SBI could estimate model parameters based on human behavior data and qualitatively reproduce human behavior (i.e., SBI-estimated F1 trajectories vs. human behavioral data). Using the same $p(\theta|\boldsymbol{x})$ estimated in the first assessment, we instead used the mean of the healthy speakers'

online compensation (Parrell et al., 2017) as our observation, and estimated joint and marginal distributions $p(\theta|\boldsymbol{x}_{\text{Healthy}})$. The modes of the distributions were used to generate SBI-estimated-F1 trajectories via FACTS. This SBI-estimated F1 trajectories were compared with the behavioral data (i.e., group mean).

Results. In our simulated dataset, SBI was able to estimate the ground-truth FACTS parameters used to generate the data (Figure 1A). Using these parameters to generate new model data with FACTS resulted in a close fit with the behavioral ground-truth data used for model inference (Figure 1B). SBI-derived estimates of model parameters for the real compensatory behavior data from neurobiologically healthy speakers resulted in joint distributions that were mostly gaussian with few showing correlating or skewed structure (Figure 1C). As for the model data, using the SBI-derived parameter values resulted in a compensatory response qualitatively similar to the behavioral data used to train the model (Figure 1D).

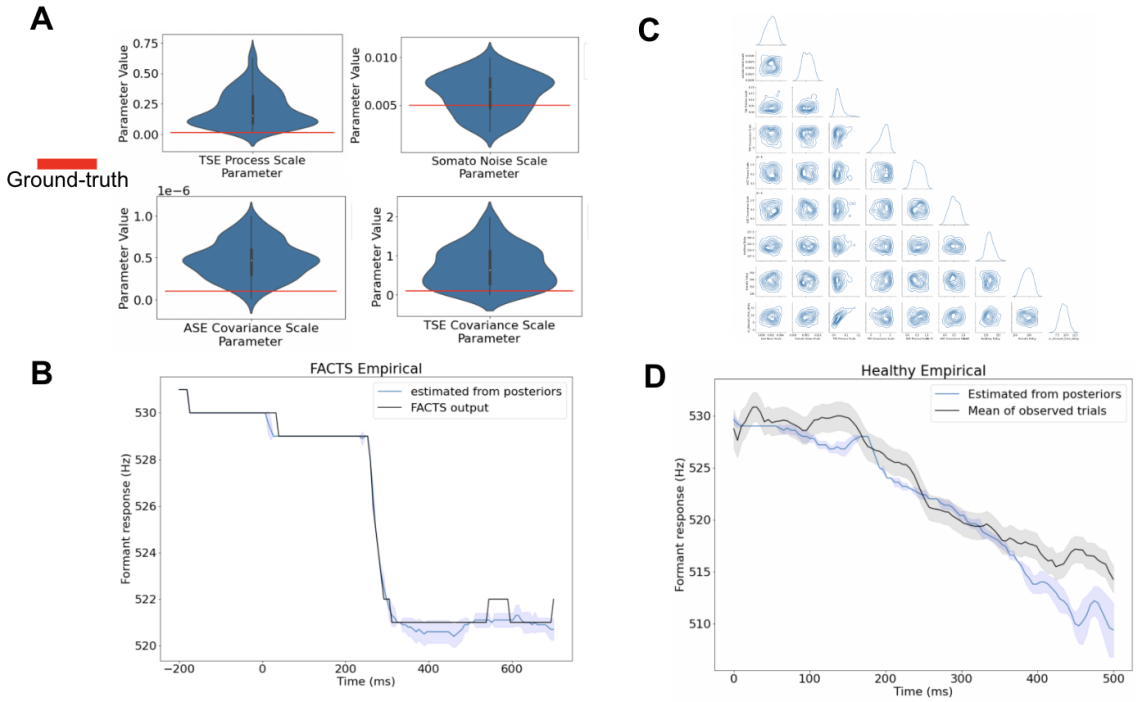


Figure 1: *Model inference results.* ***A****: Violin plots showing select parameters' posterior distributions. The red lines show ground-truth values.* ***B*** *Line plots comparing a FACTS empirical F1 trajectory to SBI-estimated F1 trajectories.* ***C****: Joint and marginal posterior distributions derived from the healthy empirical F1 dataset.* ***D****: Line plots comparing the Healthy Empirical F1 trajectories to SBI-estimated F1 trajectories.*

Discussion. This work validates the ability of SBI to recover a known parameter set in the FACTS model, which is a necessary first step towards its use in estimating changes in control in speakers with neurogenic speech disorders. We additionally showed that SBI-derived model parameters can provide a good qualitative fit to human behavioral data. These validation assessments motivate future work using SBI with FACTS on behavioral data from patients with neurogenic speech disorders, such as ataxic dysarthria, to uncover the underlying neural computational differences that explain changes in behavior between healthy and disordered populations. The unique benefit of the modeling approach lies in its ability to offer mechanistic insights into the neural processes underlying speech control and compensation which cannot be directly inferred from behavioral data alone. Although empirical or correlative approaches may describe observed differences between groups, modeling allows researchers to propose and test hypotheses about internal computations, quantify certainty and effect sizes of each factor, thereby offering insight into what drives these differences. Such knowledge can ultimately contribute to the development of targeted interventions and therapies for speech disorders.

References

Cranmer, Kyle, Johann Brehmer, and Gilles Louppe. "The frontier of simulation-based inference." Proceedings of the National Academy of Sciences 117.48 (2020): 30055-30062.

Greenberg, David, Marcel Nonnenmacher, and Jakob Macke. "Automatic posterior transformation for likelihood-free inference." International Conference on Machine Learning. PMLR, 2019.

Guenther, Frank H., and Tony Vladusich. "A neural theory of speech acquisition and production." Journal of neurolinguistics 25.5 (2012): 408-422.

Kim, Kwang S., et al. "Mechanisms of sensorimotor adaptation in a hierarchical state feedback control model of speech." PLoS Computational Biology 19.7 (2023): e1011244.

Parrell, Benjamin, et al. "The FACTS model of speech motor control: Fusing state estimation and task-based control." PLoS computational biology 15.9 (2019): e1007321.

Parrell, Benjamin, et al. "Differential effects of cerebellar degeneration on feedforward versus feedback control across speech and reaching movements." Journal of Neuroscience 41.42 (2021): 8779-8789.

Purcell DW, Munhall KG. Compensation following real-time manipulation of formants in isolated vowels. J Acoust Soc Am. 2006; 119(4):2288–97. https://doi.org/10.1121/1.2173514 PMID: 16642842

Tejero-Cantero, Alvaro, et al. "SBI--A toolkit for simulation-based inference." arXiv preprint arXiv:2007.09114 (2020).

Model simulations suggest that speech motor control is more sensitive to estimated than true sensory noise levels

Jessica L. Gaines[1], Kwang S. Kim[2], Ben Parrell[3], Vikram Ramanarayanan[4,5], Alvincé L. Pongos[1], Srikantan S. Nagarajan[4], John F. Houde[4]

[1]UC Berkeley – UCSF Graduate Program in Bioengineering, University of California-San Francisco, San Francisco, California, USA
[2]Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, Indiana, USA
[3]Department of Communication Sciences and Disorders, University of Wisconsin-Madison, Madison, Wisconsin, USA
[4]Department of Otolaryngology, University of California-San Francisco, San Francisco, California, USA
[5]Modality.ai, San Francisco, California, USA

jessica.gaines@berkeley.edu, kwangkim@purdue.edu, bparrell@wisc.edu, vikram.ramanarayanan@modality.ai, alvince_pongos@berkeley.edu, srikantan.nagarajan@ucsf.edu, john.houde@ucsf.edu

Introduction. Participants' response to a perceived perturbation of fundamental frequency (f_0) has been used to study the effects of various neurological conditions on the speech motor control system (e.g., Houde et al., 2019). However, it can be difficult to ascribe differences in behavioral f_0 perturbation response to a particular neural mechanism. To address this, recent work has used simulation-based Bayesian inference to fit the parameters of a state feedback control (SFC) model of fundamental frequency control to observed behavioral responses to f_0 perturbations in individuals with cerebellar ataxia (CA; Gaines et al., in prep). One parameter with large effect size between the CA group and the control group is the feedback noise ratio parameter, a measurement of the relative amount of noise (as measured by the variance of the Gaussian distribution from which the noise is sampled) between two sensory modalities: auditory and somatosensory feedback (Gaines et al., in prep).

Sensory feedback noise impacts the output of the SFC model in two ways. First, Gaussian noise is added to each feedback signal at each time step. Additionally, the system estimates the amount of noise in each feedback signal and uses this to calculate Kalman gain, the scaling factor that adjusts the weight of each feedback signal in correcting the internal estimate of laryngeal state (Crevecoeur et al., 2016). Thus the large effect size of the feedback noise ratio parameter could suggest that the two groups differ in relative feedback noise between sensory modalities, or that the two groups differ in their *internal estimate* of sensory noise and calculation of Kalman gain. Here, we investigate these alternatives to better understand how the neural control of f_0 is affected by CA. Additionally, this investigation will provide more detail on the role of feedback noise in the SFC model.

Methods. In this state feedback model of f_0 control (Houde & Nagarajan, 2011; Houde et al., 2014), a controller generates motor commands based on an internal estimate of the laryngeal state (position and velocity). An efference copy of these commands is used to predict the subsequent laryngeal state, and thus the expected sensory consequences of the movement. The plant, a simplified model of the larynx (a state space representation of a spring-mass system) generates auditory and somatosensory feedback. This simulated feedback, which is delayed and combined with Gaussian noise, is compared with the expected sensory feedback, and then the error between these two signals is used to correct the internal estimate of laryngeal state. This correction is weighted by a Kalman gain that scales the weight of each error signal based on the amount of noise in the feedback. This function that predicts sensory feedback and corrects the internal state estimate is referred to as the observer. The observer was previously assumed to ideally estimate sensory feedback noise, that is, an exact copy of the variance of sensory feedback noise was available for use by the observer. For this investigation, the actual feedback noise in the plant was separated from the observer's estimate of feedback noise, so that these two parameters were independently tunable.

A parameter sweep of auditory and somatosensory feedback noise variance (the variance of the Gaussian distribution from which the noise was sampled) was conducted for each of these conditions. The baseline model output was set to closely match the behavioral pitch perturbation response of the control group from Houde et al. (2019) by using the inferred parameter set from Gaines et al. (in prep). The inferred value of auditory noise variance was 1.5e-6 and the inferred value of somatosensory noise variance was 7.3e–7. Four parameter sweeps were then conducted. First auditory noise variance was swept twice across the following values while somatosensory noise variance was held constant at the inferred value: [1e-6, 1.2e-6, 1.5e-6, 2e-6, 3e-6]. During the first sweep, the true noise in the plant was held constant at the inferred value while the internal estimate of feedback noise was swept. During the second sweep, the true noise in the plant was swept while the estimate of noise in the observer used to calculate the Kalman gain was held constant at the inferred value. Next, somatosensory noise variance was swept for each of these conditions while auditory noise was held constant at the inferred value. Somatosensory noise was swept across the following values: [5e-7, 6e-7, 7.3e-7, 1e-6, 2e-6].

Results. As seen in **Figure 1**, changes in the internal estimate of auditory and somatosensory noise cause drastic changes in model output. However, when noise parameters are isolated from the Kalman gain and only the true noise in the plant is allowed to vary, the noise parameters have negligible change on the simulated pitch response, with indistinguishable model outputs across the parameter sweep.

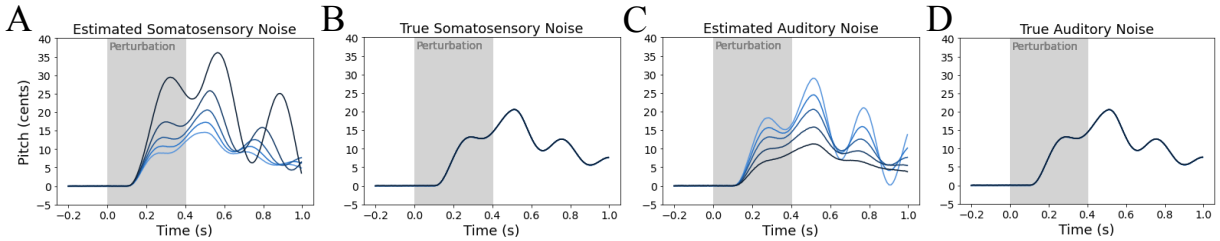


Figure 1: A,C) Changes in the internal estimate of feedback noise alone cause changes in model output. B,D) Changes in the actual feedback noise alone cause no changes in model output.

Discussion. The results indicate that the importance of the sensory feedback noise parameters in the SFC model lies entirely in their role in the calculation of Kalman gain. This impacts model output because the relative feedback noise between sensory modalities is used to weight each sensory signal in updating the internal state estimate. For tasks in which the perturbation is in the auditory domain only, greater auditory noise relative to somatosensory noise decreases the weighting of the auditory feedback signal and reduces the magnitude of the response, while greater somatosensory noise decreases the relative weighting of the unaltered somatosensory signal and increases the magnitude of the response. Thus while differences in true feedback noise may exist between the CA group and the control group, these results suggest that the more critical difference is the internal estimate of feedback noise, which may or may not relate to actual changes in sensory reliability (e.g., estimates of sensory delay may impact Kalman weights; Crevecour et al., 2016).

References

Crevecoeur, F., Munoz, D. P., & Scott, S. H. (2016). Dynamic Multisensory Integration: Somatosensory Speed Trumps Visual Accuracy during Feedback Control. Journal of Neuroscience, 36(33), 8598–8611. doi: 10.1523/jneurosci.0184-16.2016

Gaines, J.L., Kim, K.S., Parrell, B., Ramanarayanan, V., Pongos, A.L., Nagarajan, S.S., & Houde, J.F. (in prep.). Bayesian inference of state feedback control parameters for f_0 perturbation responses in cerebellar ataxia.

Houde, J.F. & Nagarajan, S.S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience 5*, Article 82. doi: 10.3389/fnhum.2011.00082

Houde, J.F., Niziolek, C.A., Kort, N., Agnew, Z., & Nagarajan, S.S. (2014, May 5-8). Simulating a state feedback model of speaking. 10th International Seminar on Speech Production, Cologne, Germany.

Houde, J.F., Gill, J.S., Agnew, Z., Kothare, H., Hickok, G., Parrell, B., et al. (2019). Abnormally increased vocal responses to pitch feedback perturbations in patients with cerebellar degeneration. *Journal of the Acoustical Society of America 145*(5), EL372-EL378. doi: 10.1121/1.51009

An automated pipeline for preprocessing spontaneous L2 English prosody

Sylvain Coulange[1,2,3], Tsuneo Kato[3], Solange Rossato[2], Monica Masperi[1]

[1] *Univ. Grenoble Alpes, Laboratory of Linguistics and Didactics of Foreign and Mother Tongues (LIDILEM), 38000 Grenoble, France*

[2] *Univ. Grenoble Alpes, CNRS, Institute of Engineering, Grenoble Computer Science Laboratory (LIG), 38000 Grenoble, France*

[3] *Doshisha Univ., Spoken Language Processing Laboratory (SLPL), 610-0394 Kyoto, Japan*

sylvain.coulange@univ-grenoble-alpes.fr, tsukato@mail.doshisha.ac.jp,
solange.rossato@univ-grenoble-alpes.fr, monica.masperi@univ-grenoble-alpes.fr

While numerous tools address L2 pronunciation, they tend to focus on segmental deviations, often neglecting prosody and lacking pedagogical feedback (Coulange 2023). In contemporary L2 speaking classes, the foremost priority is achieving "understandability," encompassing both being understood and achieving it as effortlessly as possible (commonly referred to as intelligibility and comprehensibility, Derwing and Munro 2015). Within this framework, assessing speech requires identifying phenomena that significantly hinder listener understanding, and prioritizing them in assessment. Pinpointing these target areas in students' speech facilitates focused improvement for enhanced comprehensibility.

In the realm of English as a foreign language, rhythm, notably the placement of hesitation markers like silent or filled pauses and lexical stress realization, plays a crucial role in comprehensibility. Conversely, common segmental, grammatical, and lexical deviations show a comparatively lower impact on the cognitive load associated with speech processing, though they remain important considerations (Isaacs, Trofimovich, and Foote 2018; Walker, Low, and Setter 2021; Tortel 2021, among others). While some tools analyze pause frequency and length (de Jong, Pacilly, and Heeren 2021) or classify lexical stress (Ferrer et al. 2015; Shahin, Epps, and Ahmed 2016), our investigation identified a gap in tools considering the syntactic context of pauses and the degree of contrast between stressed and unstressed syllables. We developed a fully automated pipeline for processing spontaneous L2 English speech, that analyzes pausing and stress patterns.

Two releases of the Pauses and Lexical Stress Processing Pipeline[1] (plspp) currently coexist. Both are based on WhisperX speech recognition (Bain et al. 2023), but the first one (plspp1) extracts stress related acoustic parameters from syllable nuclei points, while the second version (plspp2) uses an extra layer of phoneme-level forced alignment using Montreal Forced Aligner (McAuliffe et al. 2017) and extracts acoustic parameters within vowel intervals. Pause pattern analysis is based on inter-word intervals' duration, part-of-speech context, opening and closing constituents, considering their size and syntactic depth (Kitaev, Cao, and Klein 2019). Pauses lower and upper duration thresholds can be easily set up to consider only intervals of a certain duration.

The analysis of lexical stress involves comparing word-level prosodic shapes with their expected stress pattern extracted from a reference dictionary, and measuring the prosodic contrast between stressed and unstressed syllables. Each syllable is represented by three speaker-normalized measures: F0, intensity and duration.

[1] The pipeline is open-source and freely available here https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plspp.

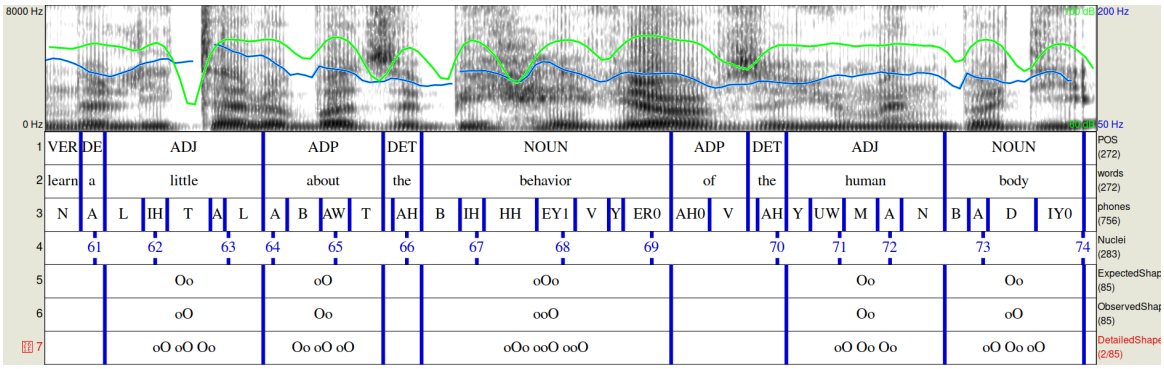


Figure 1: Example of a TextGrid output from plspp2 showing POS tags (1), transcribed text (2), phoneme alignment (3), syllable nuclei (4), expected prosodic shape (5), observed prosodic shape (6), F0, intensity and duration shapes (7)

Acoustic stress is inferred to be the most prominent syllable within the word for each dimension, and these three dimensions are merged with equal weight to obtain a single global representation easier to handle. Stress position is analyzed through a binary representation of syllables, with "O" representing the stressed syllable and "o" representing the other syllables in the word. Both releases do not consider the secondary stress yet.

Both versions output several tables including one listing the stressed words with their acoustic detailed information, and another table listing all inter-word intervals along with their duration and syntactic context for pause pattern analysis. Moreover, a TextGrid file is generated for each audio file allowing further acoustical analysis (cf. Figure 1).A visualisation tool is also being developed in order to more easily overview – and dive in – the results. This tool exists as a light standalone html/js-only version encompassed in the plspp pipeline; as well as a Django server-based application for web hosting purposes.

This pipeline has already been used in several studies involving French, Japanese and Korean learners of English, as well as native speakers of English, in elementary school and at university, on spontaneous, recited or read aloud speech.

Our presentation will describe how both pipeline work and elucidate the decision-making process behind them, thereby initiating a discussion about their inherent limitations and possible future improvements. Additionally, we will showcase the different ongoing studies, presenting preliminary results that have been obtained thus far.

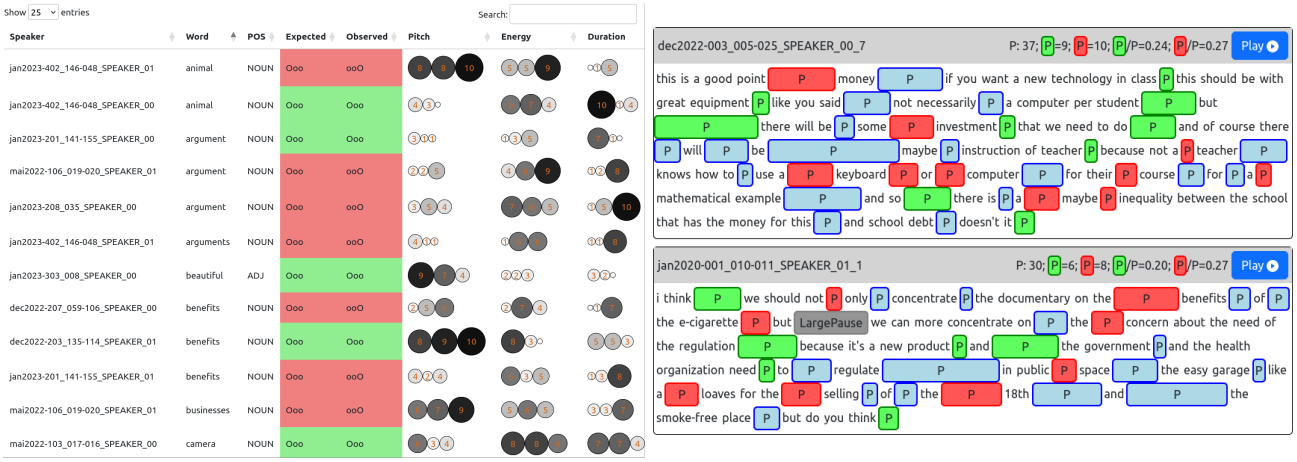


Figure 2: Overview of the stress pattern analysis (left) and pause patterns (right), with inter-clause pauses in green, inter-phrase in blue and intra-phrase in red

References.

Bain, Max, Jaesung Huh, Tengda Han, and Andrew Zisserman (2023). "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio". In: *Interspeech*.

Coulange, Sylvain (2023). "Computer-aided pronunciation training in 2022: When pedagogy struggles to catch up". In: *Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices*. Ed. by Alice Henderson and Anastazija Kirkova-Naskova, pp. 11–22. DOI: 10.5281/zenodo.8137754.

de Jong, Nivja H., Jos Pacilly, and Willemijn Heeren (2021). "PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically". In: *Assessment in Education: Principles, Policy & Practice* 28.4, pp. 456–476. DOI: 10.1080/0969594X.2021.1951162.

Derwing, Tracey M. and Murray J. Munro (2015). *Pronunciation Fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins.

Ferrer, Luciana, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, and Kristin Precoda (2015). "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems". In: *Speech Communication* 69, pp. 31–45. DOI: https://doi.org/10.1016/j.specom.2015.02.002.

Isaacs, Talia, Pavel Trofimovich, and Jennifer Ann Foote (2018). "Developing a user-oriented second language comprehensibility scale for English-medium universities". In: *Language Testing* 35.2, pp. 193–216. DOI: 10.1177/0265532217703433.

Kitaev, Nikita, Steven Cao, and Dan Klein (2019). "Multilingual Constituency Parsing with Self-Attention and Pre-Training". In: *ACL*. Florence, Italy, pp. 3499–3505.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi". In: *Proc. Interspeech 2017*, pp. 498–502. DOI: 10.21437/Interspeech.2017-1386.

Shahin, Mostafa Ali, Julien Epps, and Beena Ahmed (2016). "Automatic Classification of Lexical Stress in English and Arabic Languages Using Deep Learning". In: *Interspeech*, pp. 175–179.

Tortel, Anne (2021). "Le rythme en anglais oral : considérations théoriques et illustrations sur corpus". In: *Recherche et pratiques pédagogiques en langues - Cahiers de l'APLIUT* Vol. 40 N°1. DOI: 10.4000/apliut.8857.

Walker, Robin, Ee-Ling Low, and Jane Setter (2021). *English pronunciation for a global world*. Oxford University Press.

Discovering phoneme-specific critical articulators through a data-driven approach

Jesuraj Bandekar, Sathvik Udupa, Prasanta Kumar Ghosh

Electrical Engineering Department, Indian Institute of Science, Bangalore

jesurajbandekar.661@gmail.com, prasantg@iisc.ac.in

Introduction. Speech articulators such as tongue, lips, jaw and velum play an important role in speech production. The varying positions of these articulators result in sounds referred to as phonemes. It is known that, for different phonemes, certain articulators consistently achieve their target positions, and these are known as critical articulators [Philip J.B. Jackson et al. (2009)]. The non-critical articulators' positions tend not to affect the characteristics of the phoneme produced. Let us consider the use of time-varying articulatory trajectories obtained from Electromagnetic articulography (EMA); the presence of non-critical articulators may be a source of noise for various modelling tasks such as Acoustic to Articulatory Inversion (AAI), Articulatory to Acoustic Forward (AAF) mapping, Phoneme to Articulatory (PTA) mapping etc. Thus, we believe that measuring the degree of criticality of articulators for different phonemes is beneficial. There has been previous work by PK Anusuya et al. (2020) which measures the criticality of articulators using the data distribution of articulatory movements. In this work, we are interested in exploring if such observations can be discovered unsupervised while training a neural network model for phoneme classification task.

Methods. We propose to learn phoneme-specific critical articulators unsupervised through a data-driven end-to-end machine-learning approach. To achieve this, we use two blocks of neural networks to perform three tasks - AAI; Peng Liu et al. (2015) and Aravind Illa et al. (2018), AWP (articulator weight prediction) and FPC (frame-level phoneme classifier). These three tasks are learned end-to-end with speech features as the input and frame-level phonemes as the final output, along with time-varying articulatory features and time-varying articulatory weights as intermediary outputs. We perform AAI using a neural network, mapping input 13-dim MFCC features to 12-dim articulatory trajectories, x and y coordinates of six articulators: UL, LL, JAW, TT, TB, and TD. This 12-dim data is the intermediary output, which is trained with Mean squared error (MSE) loss with ground truth articulatory movements. For AWP, we use a similar neural network and a different dense output layer to predict 12-dim features from input MFCCs. These features are further normalised through min-max normalisation across the 12 features. This acts as normalised weights for the articulators. Finally, we multiply the normalised weights and articulatory features to weigh the articulators; this acts as the input to FPC. The FPC is another neural network which predicts frame-level phoneme labels. This is optimised using frame-level cross-entropy loss against ground truth frame-level phonemes. We use transformer-based neural architecture for all models, following the architecture used by AAI in Sathvik Udupa et al. (2021).

We use the network described above to allow the model to estimate phoneme-specific critical articulators based on the weights learnt, and we find all three tasks are necessary for this purpose. We need AWP to predict features that can act as weights for articulators, where higher weights represent critical articulators. However, AWP doesn't have an objective function; it is learned unsupervised. For it to learn higher weights for critical articulators, it needs a task which assists this learning. Hence, we use FPC so that the FPC loss can guide the AWP network to learn the required features. Now, the need for an AAI network arises - theoretically, ground truth articulatory movements could be multiplied with AWP weights for the FPC network. However, during training, we find that having an AAI network to predict articulatory movements improves the accuracy of predicted weights. We hypothesise that this is the case as it benefits phone prediction through the FPC - AAI backward pass, allowing more access to MFCC features. We add additional optimisations to achieve better weights from the AWP network. Firstly, after AAI, we use a straight-through estimator (STE) [Yoshua Bengio et al. (2013)] to replace predicted articulators with ground truth articulators while maintaining gradient flow from FPC to AAI, i.e., STE acts as an identity function for the backward pass. Additionally, we use a dropout of 0.5 on the 12-dim weight prediction features in AWP. We find that, without dropout, the AWP weights tend to get activated for particular articulators for all phonemes.

To summarise, the goal of our model formulation is to learn frame-level weightage across articulators (AWP), where high scores represent critical articulators. We then add a task, i.e., FPC, where this information could be learnt unsupervised.

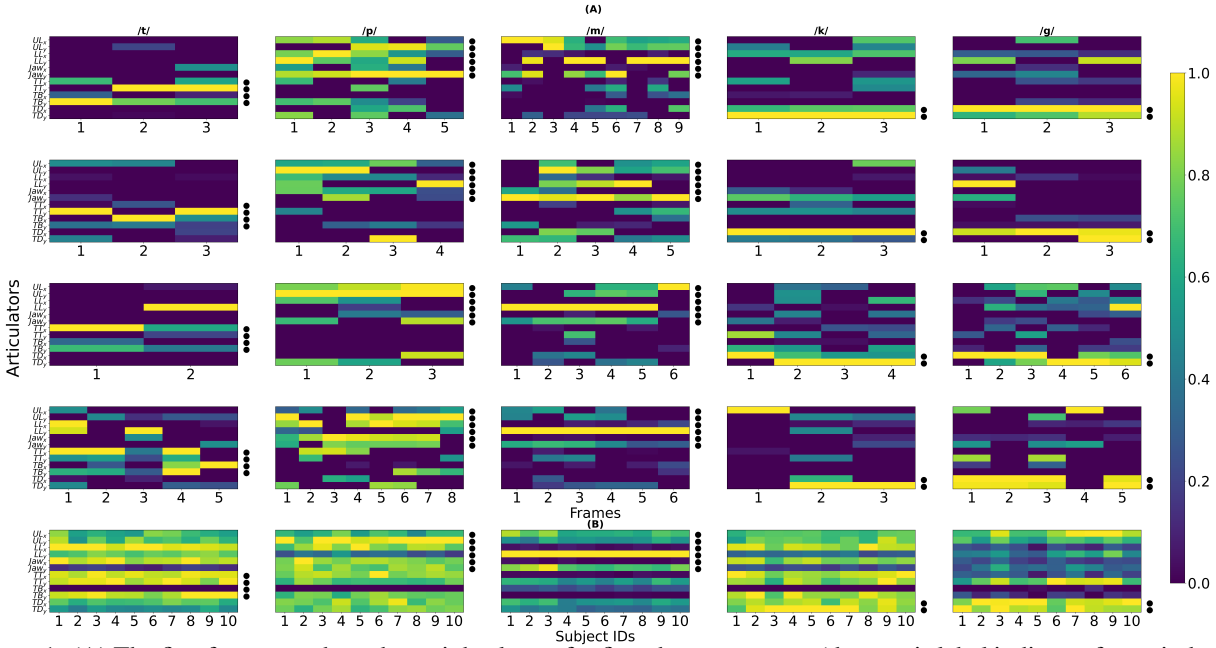


Figure 1: (A) The first four rows show the weights learnt for five phone segments (the x-axis label indicates frame index) in different utterances. The articulators known to be critical are marked by . (B) The last row represents the corresponding phoneme's average representation across all segments. It is shown for all 10 subjects (the x-axis label indicates the frame index) used in this study.

Finally, we add various optimisation methods to improve the weight predictions from AWP (AAI, dropout, STE). We implement our network in PyTorch and train the model on data of around 5 hours from 10 subjects comprising both acoustic and articulatory data. The implementation is available on our public GitHub repository[1].

Results and Discussion. Here, we analyse the effectiveness of the weights predicted for articulators. We visualise the normalised weights for five different phonemes (one phoneme in one column) for four segments as shown in Figure 1. Additionally, we also show the average representation of all phoneme segments in the last row. We can observe that the critical articulators [2] are consistently activated for the relevant phonemes. For example, for **/t/**, we can observe that many tongue articulators are prominently activated. Additionally, it can be seen that the axis of the critical articulator can also learned in this process, such as activating only x or y axis. For example, for **/m/** LL_y is activated, while LL_x is not; this is in line with speech production as the movement along y axis is necessary to form the constriction to produce **/m/**. This validates using our unsupervised approach to learn the weightage of critical articulators using an auxiliary task. Rather than grouping articulators as critical or non-critical, we get a measure of criticality from the weights. For example, for **/k/** we can observe other articulators, such as lips and tongue, are slightly activated; This could help in several tasks where even non-critical articulators can play a role. Next, we use the speaker-level data from Figure 1(B) and compute the overall phoneme level average across articulators. These are the top three articulators activated for each phoneme - **/t/** : $LL_x\ TT_y\ TT_x$, **p** : $UL_y\ TT_x\ LL_x$, **m** : $LL_y\ Jaw_y\ UL_x$, **k** : $TD_y\ TT_x\ LL_x$ and **g** : $TD_x\ TD_y\ UL_x$. We can observe that for these five phonemes, there are at least 2 critical particulars activated in the top three, except for **/k/** where a single articulator is present. We believe that other phonemes are also activated due to co-articulation.
In the future, we will look at improving the weightage of articulators and identify the effect of co-articulation. Further, we plan to use the weights for different tasks involving articulatory trajectories.

References.

Aravind Illa et al. (2018). "Low Resource Acoustic-to-articulatory Inversion Using Bi-directional Long Short Term Memory." In: *Interspeech*.

Peng Liu et al. (2015). "A deep recurrent approach for acoustic-to-articulatory inversion". In: *ICASSP*. IEEE, pp. 4450–4454.

Philip J.B. Jackson et al. (2009). "Statistical identification of articulation constraints in the production of speech". In: *Speech Communication* 51.8.

PK Anusuya et al. (2020). "A data driven phoneme-specific analysis of articulatory importance". In: *International Seminar On Speech Production*.

Sathvik Udupa et al. (2021). "Estimating Articulatory Movements in Speech Production with Transformer Networks". In: *Proc. Interspeech 2021*.

Yoshua Bengio et al. (2013). *Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation*. arXiv: 1308.3432.

[1] https://github.com/coding-phoenix-12/CriticalArticulatorUSL

[2] https://home.cc.umanitoba.ca/ krussll/phonetics/ipa/articulatory-ipa.html

Cerebellar degeneration impairs adaptation to pitch perturbations in sustained vocalization

Anneke Slis[1], Benjamin Parrell[1,2]

[1]Waisman Center, University of Wisconsin–Madison,
[2]Communication Sciences and Disorders, University of Wisconsin–Madison
slis@wisc.edu, bparrell@wisc.edu

Introduction: When exposed to external perturbations of their vocal pitch, neurobiologically healthy speakers (NHS) change their production to oppose the perturbation, using both feedback and feedforward control mechanisms. At the onset of a pitch perturbation, speakers use feedback control to oppose the perturbation within the ongoing vocalization, referred to as compensation or reflex response (Burnett et al., 1998). When a perturbation is consistently applied across consecutive productions, speakers additionally adapt their feedforward motor commands to oppose the perturbation on subsequent trials (Jones & Munhall, 2000). This sensorimotor adaptation in reaching, locomotion, and supralaryngeal articulation is thought to rely on the cerebellum, based on reduced adaptation observed in individuals with cerebellar ataxia (CA) in these tasks. Compensation, on the other hand, has been shown to be intact in CA across motor domains. Strikingly, individuals with CA not only have intact compensation in pitch control, but uniquely among motor domains have been shown to exhibit larger compensatory responses (Houde et al., 2019; Li et al., 2019). However, no studies to date have examined whether cerebellar damage impairs adaptation in vocal pitch. Critically, controlling vocal pitch differs from other motor behavior as it may be more reliant on feedback versus feedforward control. For example, pitch control in post lingually deaf individuals deteriorates much more rapidly than supralaryngeal articulation. Here, we evaluate whether and how individuals with CA differ from NHS in pitch adaptation, hypothesizing that individuals with CA show reduced adaptation compared to NHS.

Methods: 19 individuals with CA (8 M ([62-74]); 11 F ([25-70]) and 27 age-matched NHS (10 M ([39-72]); 17 F ([31-80])) participated in three sessions: one measuring pitch adaptation, one control, and one measuring compensation. All participants were native American English speakers, and reported no hearing, speech, language, or neurological impairments other than ataxia. In all sessions, participants produced a sustained vowel ("ah") for 1 s on each trial. The adaptation session consisted of a 20-trial unperturbed baseline phase, a 40-trial hold phase with a constant -100 cent pitch shift, and a 20-trial unperturbed washout phase. The control session was identical, except that no perturbation was applied. The compensation session consisted of 105 trials. On a pseudo-random subset of 80 of these productions, pitch was either shifted down (40 trials) or up (40 trials) ±100 cents, starting between 200 ms to 500 ms after vocalization onset and continuing until the end of the trial. Audapter (Cai et al., 2008) was used to record speech at 16 kHz, process (and on some trials alter) the pitch, and play back the produced speech over closed-back headphones.

For the adaptation and control sessions, pitch contours from Audapter were used to calculate the median pitch value over the first 100 ms of the vowel on each trial (avoiding any online compensatory response) and normalized using the median pitch during the 20 baseline trials of each session. Adaptation was measured from the final 20 productions of the hold phase. In addition, for each session, we calculated within-trial compensation to assess online feedback corrections over and above any potential adaptation. Compensation was measured as the median pitch difference between vowel onset (0-100ms) and a later window from 300-400 ms after vowel onset, using data from all 40 productions in the hold phase. For the compensation task, compensation was similarly taken as the change in pitch in the window from 300-400 ms after perturbation onset, normalized for each trial based on the median pitch during a 100ms window immediately prior to the onset of the pitch perturbation, matching methods used in previous work (Houde et al., 2019, Li et al., 2019).

For adaptation and online compensation in the adaptation task, linear mixed models were created (lme4 package, Bates et al., 2015), with factors of session (adaptation [perturbation applied during hold phase], control [no perturbation]), group (CA, NHS), their interaction, and random intercepts for participants. For the compensation analysis, a similar model was used, with session replaced with the factor condition (upward perturbed, downward perturbed, unperturbed). Statistical significance was assessed with the lmerTest package (Kuznetsova et al., 2017). Pairwise comparisons were performed using EMMEANS (Lenth et al., 2022). Data for each analysis were transformed to correct for non-normal distributions using an Ordered Quantile normalization transformation (Peterson & Cavanaugh, 2020).

Results: For adaptation, we observed a significant interaction between group and session ($\beta = -0.28$, $t(1794) = -3.42$, $p < 0.001$), indicating that the two groups differed in their adaptive response to the auditory perturbation (**Figure 1A**). Only NHS showed increased pitch in the adaptation session relative to the control session (31±5 cents, $t(1796) = 5.99$, $p < 0.0001$). Conversely, individuals with CA did not show any difference between the two sessions (3±7 cents, $t(1796) = -0.42$, $p = 0.67$). Over and above any feedforward changes in the adaptation task, NHS compensated substantially for the onset perturbation in the adaptation session (13 ± 3 cents, $z = 4.35$, $p < .0001$). Conversely, the pitch of CA declined in

the perturbed session (-14±4 cents, z = -3.89, p < .001), leading to a main effect of session (β = 0.14, t(3634) = 3.89, p < 0.001) and an interaction between group and session (β = -0.27, t(3634) = -5.75, p < .0001, **Figure 1B**).

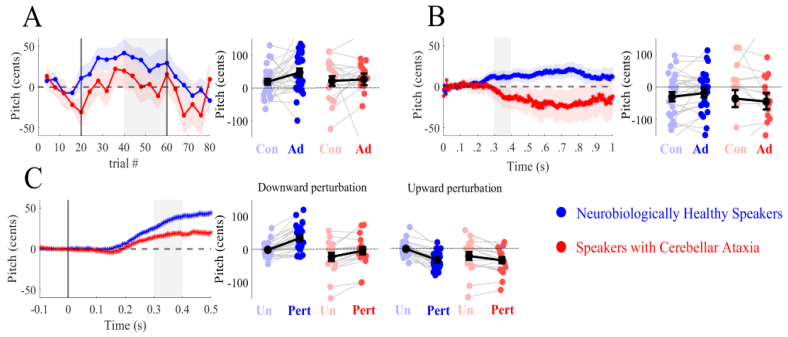


Figure 1. ***A****: Adaptation. Left: Pitch changes across experiment (adaptation minus control sessions); right: individual data for control (Con) and adaptation (Ad) sessions.* ***B****: Online compensation to onset perturbation. Left: pitch tracks over time (adaptation minus control sessions); right: individual data as in A.* ***C****: Online compensation to mid-utterance perturbation in compensation session. Left: pitch tracks over time (perturbed minus unperturbed conditions, upward perturbation response sign-flipped); right: individual data for unperturbed (Un) and perturbed (Pert) trials.*

During the compensation session (**Figure 1C**), both CA and NHS opposed the perturbation. For CA, both responses to downward (24±4 cents, z = 6.55, p < .0001) and upward perturbations (12±4 cents, z = 3.22, p < .01) differed from unperturbed trials. For NHS, these responses were larger, with responses to both downward (38±3 cents, z = 12.69, p <.0001) and upward perturbations (31±3 cents, z = 10.50, p <.0001) differing from unperturbed trials. The larger response in NHS led to significant interactions between group and downward perturbed (β = 13.80, t(4637) = 2.96, p < 0.01) and group and upward perturbed (β = -19.42, t(4637) = -4.17, p < 0.0001).

Discussion: Our results show that adaptation of feedforward control of vocal pitch in response to external pitch perturbations was substantially impaired (essentially eliminated) in individuals with CA. This suggests that adaptation in pitch, like other motor domains, relies on a cerebellar-dependent learning mechanism for updating motor plans. These results suggest that cerebellar degeneration causes an even more dramatic impairment in the adaptation of feedforward motor plans in pitch than in other domains, including vowel formant control, where these patients show impaired, but present, adaptive responses. Surprisingly, online compensation to mid-utterance perturbations during sustained vocalisations was present but reduced in magnitude relative to controls. These results are at odds with previous studies (Houde et al., 2019; Li et al., 2019), both of which show significantly *enhanced* compensatory responses to mid-utterance perturbations of vocal pitch in individuals with CA relative to NHS. It is possible that methodological differences may underlie these distinct findings: in both previous studies, pitch perturbations were transiently delivered on every trial. It is possible that these frequent perturbations cause individuals with CA to become more reliant on feedback for pitch control than they otherwise would be. Conversely, the differences may be due to random differences in sampling. Consistent with this latter idea, the magnitude of upward compensation to mid-utterance perturbations in our sample of individuals with CA (~24 cents) is similar to previously reported values (~30 cents in Houde et al.), but the response we observed in NHS was substantially larger than in previous studies (~38 cents vs ~12 cents in Houde et al., 2019). Strikingly, however, we found that online compensatory responses to perturbations at vowel onset in the CA group were completely absent. This dissociation between responses to mid-utterance and onset perturbations suggests that these may rely on different mechanisms, with the latter potentially tapping into mechanisms to maintain a stable pitch while the former may entail comparisons with sensory predictions that are likely impaired in the CA group.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1–48.

Burnett, T. A., Freedland, M. B., Larson, C. R., Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America,103*(6), 3153-61.

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2008). A System for Online Dynamic Perturbation of Formant Trajectories and Results from Perturbations of the Mandarin Triphthong /iau/. *Proceedings of the 8th International Seminar on Speech Production,* 65–68.

Houde, J. F., Gill, J. S., Agnew, Z., Kothare, H., Hickok, G., Parrell, B., Ivry, R. B., & Nagarajan, S. S. (2019). Abnormally increased vocal responses to pitch feedback perturbations in patients with cerebellar degeneration. *The Journal of the Acoustical Society of America, 145*(5):EL372.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America, 108*(3), 1246-1251.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(1), 1-26.

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (1.8.2,) [Computer software]. https://cran.r-project.org/web/packages/emmeans/index.html.

Li, W., Zhuang, J., Guo, Z., Jones, J. A., Xu, Z., & Liu, H. (2019). "Cerebellar Contribution to Auditory Feedback Control of Speech Production: Evidence from Patients with Spinocerebellar Ataxia." *Human Brain Mapping 40*(16), 4748–58.

Peterson, R. A., & Cavanaugh, J. E. (2020). "Ordered quantile normalization: a semiparametric transformation built for the cross-validation era." Journal of Applied Statistics, 47(13-15), 2312-2327.

Speech sensorimotor adaptation in individuals with hearing-impairment

Monica Ashokumar[1], Jean-Luc Schwartz[1], Takayuki Ito[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France
monica.ashokumar@grenoble-inp.fr,takayuki.ito@gipsa-lab.grenoble-inp.fr

Introduction. The importance of auditory feedback for learning and maintaining speech production has been frequently examined using an experimental adaptation model based on real-time formant modulation of altered auditory feedback (Houde & Jordan, 1998; Villacorta et al., 2007). When a produced vowel sound is played back with altered formants during a vowel repetition task, speakers adapt their production in the direction opposite to the imposed formant change which gradually returns to original after the removal formant perturbation as an aftereffect. Although this is consistent in normal-hearing individuals and in some disorders such as blind speakers (Trudeau-Fisette et al., 2017), Parkinson's disease (Mollaei et al 2013) and person who stutter (Daliri et al 2018), it is still unknown whether hearing-impaired individuals using hearing devices also show adaptation to real-time formant modulation. Previous studies have shown that the cochlear-implanted individuals can adapt to other types of altered feedback such as delayed auditory feedback and frequency altered feedback (Taitelbaum-Swead et al., 2019) suggesting that such participants could also adapt to formant perturbations. We examined whether hearing-impaired participants using hearing devices show speech sensorimotor adaptation and how this is related to their speech perception ability and hearing profile.

Methods. We tested fourteen normal-hearing (NH) and fourteen hearing-impaired (HI) participants using hearing-aids or cochlear implant users for a long time (>10 years). All participants were French native speakers and participated in a within-subjects procedure involving speech motor adaptation test using an altered auditory feedback and speech perception tests consisting of a speech-in noise test and vowel identification test. The HI participants were asked to fill a hearing assessment questionnaire with information regarding the onset of deafness, the hearing devices used and their duration of usage. The main sensorimotor adaptation test was carried out using altered auditory feedback. We focused on the vowel /ø/. In the test, the participants were asked to repeat /ø/ in the form of the French word 'deux' (/dø/) for 150 times. The second formant in the produced sound was changed over the course of the training using Audapter (Cai et al., 2008). The first 20 trials were a baseline phase with unaltered feedback. The next 50 trials were the ramp phase where the second formant was gradually increased every trial. The next 50 trials were the hold phase with a maximum formant shift, amounting to 25% relative to the participant's second formant. The last 30 trials were an aftereffect phase where the altered feedback was removed and participants heard their own unaltered feedback. To reduce the effect of small environmental noise and bone conducted sound, small amplitude of masking noise (70 dB) was applied during the test. We evaluated the second formant, that was normalized by dividing by the mean value over the last 10 trials of the baseline phase with unaltered feedback. We quantified the adaptation amount by averaging the normalized amplitude at the end of the hold phase (average of the last 10 trials) and at the beginning of the aftereffect phase (average of the first 10 trials). Repeated-measures ANOVA was applied in each of the groups. In the speech-in-noise test, the participants identified digits presented in background noise. The signal-to-noise ratio was adjusted every trial based on their responses. The vowel identification test exploited an acoustic continuum between vowels /e/ and /ø/ including the target vowel used in the main speech motor adaptation task, the task being to categorize the presented sound as either /e/ or /ø/. We also compared how these measures concerning hearing ability and vowel identification are related to the adaptation to altered auditory feedback in HI participants.

Results. In the NH participants, the hearing thresholds in the speech-in-noise test were -9.2±0.05 dB SNR. This is close to the thresholds in a previous study (-10.5±0.3 dB SNR) in French participants (Jansen et al., 2010). The HI participants showed a varied performance in both the speech-in-noise test and the vowel identification test. Based on their hearing profile, we divided the participants into two groups. The first group (HI-CI: nine participants) were long-term cochlear implant users (≥13 years) and the second group (HI-HA: five participants) were long-term hearing-aids users (> 16 years). The first group showed hearing thresholds in noise higher than the ones by NH participants (-2.5±0.9 dB SNR), but they were able to identify the vowels in the vowel identification test. In contrast, the second group showed even higher thresholds (1.1±0.9 dB SNR) and had difficulty to identify the vowels correctly. Since the adaptation task using altered auditory feedback requires a detection of small difference between the intended produced sound and actually heard sounds, HI-HA participants may not dispose of the hearing ability required for adaptation, contrary to HI-CI ones. This is confirmed by experimental data displayed in Figure 1, which shows the averaged second formants in the baseline phase, the hold phase and the after-effect phase. One-way repeated measures ANOVA showed that the second formant was significantly different across these three phases in NH ($F(2, 26) = 54.9$, $p < 0.001$) and in HI-CI ($F(2, 16) = 5.3$, $p < 0.05$) but not in HI-HA ($F(2, 8) = 0.3$, $p = 0.7$). Post-hoc analyses with Tukey's HSD showed that the second formant showed

a significant decrease in the hold phase when compared with the baseline phase in NH ($z = -10.1$, $p < 0.001$) and in HI-CI ($z = -3.2$, $p < 0.01$). However, the formant was significantly different between the hold and aftereffect phase in NH ($z = -2.5$, $p < 0.05$), but not in HI-CI ($z = -2.3$, $p = 0.07$). The formant in aftereffect was also significantly different from the one in the baseline in NH ($z = 7.5$, $p < 0.001$), but not in HI-CI ($z = 0.9$, $p = 1$). Given a difference in the averaged standard deviation during the aftereffect phase between NH (0.05±0.01 (SE)) and HI-CI (0.08±0.01(SE)), the difference in aftereffect between two groups can be due to more variation in HI-CI. Individual one-sided t-tests between baseline and hold phase were carried out to verify if the adaptation was induced in individual-bases. In this test, 6 of 9 HI participants and 13 of 14 NH participants showed a significant decrease. Although the rate of 66% of participants showing an adaptation is comparable with the previous finding of adaptation (Lametti et al., 2012), this was smaller than the one in NH participants (93%) in the current data. This smaller adaptation ratio in HI participants can result in the relatively smaller amplitude of the adaptation in the averaged data. This could also be due to the unequal size of the groups (14 NH vs 9 HI-CI).

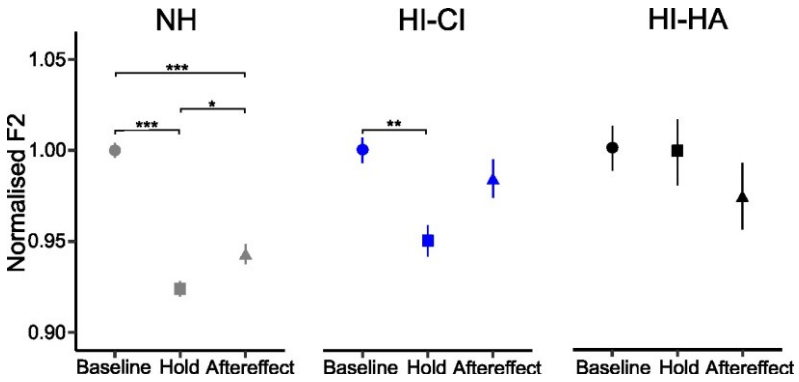


Figure 1. *Averaged formants in baseline, hold and aftereffect phase. The left panel corresponds to NH group, middle is HI-CI group and the right is HI-HA group. The error bars represent standard errors across the participants. *: p<0.05, **: p< 0.01, and ***: p<0.001.*

Discussion. This study assessed whether HI participants using hearing devices show speech sensorimotor adaptation and whether this is dependent on factors affecting their hearing. We confirmed that our experimental setup with altered auditory feedback induces an adaption in NH participants as shown in previous studies (Trudeau-Fisette et al., 2017). Using this setup, we found that HI participants showed an adaptation only in the HI-CI group but not in the HI-HA group. The nature of auditory input from these two devices can be different because hearing-aids simply provide amplification to the sounds entering the ear in comparison to cochlear implants which bypass the damaged ear regions and directly stimulate the auditory nerve. This difference could also be seen in the result of hearing threshold in noise and vowel identification performance. In addition, while some participants in the HI-HA group have used cochlear implant for a short period, they still did not show any adaptation behavior. Altogether this suggests that HI-HA participants have a reduced ability to precisely distinguish between the minute speech sound differences compared with HI-CI ones and that a certain period of cochlear implant experience maybe required to achieve this. These findings highlight the importance of auditory feedback in controlling production mechanisms and that in HI individuals such control of speech production may be related to factors affecting their hearing.

References

Cai S, Boucek M, Ghosh SS, Guenther FH, Perkell JS. (2008). A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/. In Proceedings of the 8th Intl. Seminar on Speech Production, Strasbourg, France, Dec. 8 - 12, 2008. pp. 65-68.

Daliri A, Wieland EA, Cai S, Guenther FH, Chang S-E (2018) Auditory-motor adaptation is reduced in adults who stutter but not in children who stutter. Dev Sci 21.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor Adaptation in Speech Production. *Science*, *279*(5354), 1213–1216. https://doi.org/10.1126/science.279.5354.1213

Jansen, S., Luts, H., Wagener, K. C., Frachet, B., & Wouters, J. (2010). The French digit triplet test: A hearing screening tool for speech intelligibility in noise. *International Journal of Audiology*, *49*(5), 378–387. https://doi.org/10.3109/14992020903431272

Lametti, D. R., Nasir, S. M., & Ostry, D. J. (2012). Sensory Preference in Speech Production Revealed by Simultaneous Alteration of Auditory and Somatosensory Feedback. *Journal of Neuroscience*, *32*(27), 9351–9358. https://doi.org/10.1523/JNEUROSCI.0404-12.2012

Taitelbaum-Swead, R., Avivi, M., Gueta, B., & Fostick, L. (2019). The effect of delayed auditory feedback (DAF) and frequency altered feedback (FAF) on speech production: Cochlear implanted versus normal hearing individuals. *Clinical Linguistics & Phonetics*, *33*(7), 628–640. https://doi.org/10.1080/02699206.2019.1574313

Trudeau-Fisette, P., Tiede, M., & Ménard, L. (2017). Compensations to auditory feedback perturbations in congenitally blind and sighted speakers: Acoustic and articulatory data. *PLOS ONE*, *12*(7), e0180300. https://doi.org/10.1371/journal.pone.0180300

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America*, *122*(4), 2306–2319. https://doi.org/10.1121/1.2773966

Influence of stress and sequence position on vowel sandhi in Brazilian Portuguese

João Paulo Moraes Lima dos Santos[1]

[1]Universidad de Salamanca, Spain

joaopaulomls@gmail.com

Introduction. As in many languages around the world, Brazilian Portuguese tends to reduce vowel sequences that originally belonged to different syllables (Collischonn, 2001; Bisol, 2003), resulting in a process known as sandhi. This reduction can manifest as a monophthong (e.g., "*camisa usada*" pronounced as [kã.mi.zu.za.da]) or a diphthong ([kã.mi.zau̯.za.da]). Stress appears to be a crucial factor influencing the execution of these processes.

In the context of stress, prior studies (Tenani, 2002; Bisol, 2003) propose that sequences with stressed vowels across word boundaries are more likely to maintain hiatus, while contexts with unstressed vowels tend to undergo a sandhi process. Notably, some studies also consider the main stressed accent of the intonational phrase as a variable influencing this process (Abaurre, 1996; Bisol, 2003, 2013; Tenani, 2004). In the context of prosody, we refer to Nespor & Vogel (1986); they classify the intonational phrase (IP) as a prosodic constituent above the word, delineated by the intonational contour and pauses in speech (interpausal). According to these authors, the IP is a prosodic unit with a distinct intonation pattern and a prominent stress in its nucleus. This accentual nucleus is associated with a specific nuclear pitch that marks emphasized or new information in the sentence.

The model proposed by Nespor & Vogel (1986) posits that the IP can be subdivided into smaller domains, such as the word and the syllable, organized around a stressed accent. This implies that each prosodic domain has a primary stressed accent that weakens in a larger domain, becoming secondary. Consequently, if the vowels in a sequence receive the stressed word accent but lack it in the IP, the likelihood of applying some form of sandhi increases. Nespor & Vogel (1986) thus argue that the IP is a central prosodic domain in the organization of speech, comprising a series of tonal elements organized around an accentual nucleus.

Oliveira & Santos (2018) demonstrate this by describing the transition from the primary stressed accent at the word level (e.g., "*isso*" [ˈi.sʊ] in Portuguese) to secondary at the IP level (e.g., "*mas é isso aqui*" [i.su̯a.ˈ ki]). This shift occurs because the stressed vowel weakens when confronted with a stronger one in the IP domain, where, in Portuguese, the stressed segment consistently leans further to the right in speech. In addition to stress, we also examine the behavior of sequences based on the primary stress at the IP level along with the position of the vowel sequence. In other words, we also investigate whether the position of the sequence in the intonational phrase (within or at the pause limit of the IP) affects the vowel sequences.

Therefore, this study aims to analyze the effects of word/IP stress and vowel sequence position on sandhi processes/hiatus maintenance. The investigation explores how these factors interact and contribute to the observed sandhi and hiatus patterns in our data.

Methods. The description and analyses are based on semi-spontaneous data obtained through interviews with 10 native speakers from the city of Recife, Brazil. Recordings were conducted using a computer, a unidirectional microphone, and a Scarlett 2.0 audio interface. The data were recorded in Audacity software at a sampling frequency of 44,100 Hz. The interviews took place in offices with good acoustics, located at the Faculty of Philology or the Center for Brazilian Studies at the University of Salamanca, Spain. A total of 1,509 vowel sequences across word boundaries were obtained.

The acoustic analysis was conducted using the free software Praat (Boersma & Weenink, 2019), version 6.0.53. For statistical analysis, a mixed logistic regression model (Faraway, 2016) was employed in RStudio. This model considered three aspects: (1) the type of production as a response variable (with 1 indicating sandhi and 0 denoting hiatus maintenance in the binary relationship of the model); (2) the type of accent on vowels in the sequences (whether the vowels are stressed/unstressed), the intonational phrase stress (if the sequence receives the primary stress), and the position of the sequences (i.e., whether it is at the limit of the pause or not) as predictor variables; and (3) the speaker as a random effect. In R, the glmer function from the lme4 package (Bates et al., 2015) was employed. For a better understanding of the data in the model, log-odds results were also converted into probabilities using the invlogit function from the scales package (Wickham & Seidel, 2022) and tab_model from the sjPlot package (Lüdecke, 2018).

Results. Word stress may affect the occurrence of a sandhi process or the maintenance of the hiatus, but that will also depend on whether the vowel sequence carries the primary stress of the intonational phrase (as, for example, in "*#então no próxim***o** **A***no#*") and whether it is at the limit of the pause at the right (as in "*#pois isso qu***e** **É**#"). The results of the regression model indicate that the tendency for a sandhi to occur is higher in contexts in which at least the first vowel is unstressed, both vowels do not receive the primary stressed accent of the IP, and both vowels are not at the limit of the

pause. The intercept with a positive value (Table 1) confirms a greater probability of a contraction and a lower probability of a hiatus. However, if the context of the unstressed vowels is at the limit of the pause, the medium values reveal the hiatus preference. The negative value in log-odds (= -0.73) points to the hiatus trend, but the model cannot predict such a trend because the confidence interval values cross 0 (it can also be verified in percentages, with confidence intervals crossing 50%, or with the value p = 0.145).

In the vowel combinations V'V, the hiatus is preserved when the vowels are at the limit of the pause and carry the main accent of the IP. On the other hand, the tendency is to contract the sequence when it is in another position and does not have the primary stressed accent of the phrase.

Finally, 'VV and 'V'V vowel environments have a much higher probability of maintaining the hiatus, regardless of sequence position or the primary stress at the IP level. In both contexts, the probabilities and confidence intervals exceed 50%. A greater preference for maintaining the hiatus is also observed when these sequences receive the main stressed accent and are placed at the limit of the pause.

Sandhi			
Predictors	**Log-odds**	**IC**	***p***
(Intercept)	0.99	0.63 – 1.34	**<0.001**
word_stress [V'V]	-0.42	-0.68 – -0.17	**0.001**
word_stress ['VV]	-2.52	-3.05 – -1.99	**<0.001**
word_stress ['V'V]	-2.24	-2.90 – -1.58	**<0.001**
IP [yes]	-2.30	-2.97 – -1.61	**<0.001**
pause_limit [yes]	-0.73	-1.71 – 0.25	0.145

Table 1: *Mixed logistic regression model for the analysis of stress and vowel sequence position.*

Discussion. Our study provides empirical support for the notion that unstressed vowel contexts tend to undergo sandhi, a phenomenon influenced significantly by stress, as extensively discussed in Bisol (2003) and Tenani (2004). However, our results reveal an interesting twist: sequences with at least one stressed vowel can show a resistance to contraction. This resistance depends on how it correlates with the primary stressed accent in the IP – aligning with the studies of Abaurre (1996), Bisol (2002, 2013), and Tenani (2004) – and where the sequence sits within the utterance. Essentially, if a stressed syllable in the sequence does not match the primary stress of the IP and if the sequence is not at the limit of the pause at the right, it triggers a contraction process at the vowel boundary. This sheds light on the nuanced interplay between stress and sandhi, adding a layer of complexity to our understanding of these phonological processes in Brazilian Portuguese.

References

Abaurre, M. B. M. (1996). Acento frasal e processos fonológicos segmentais. *Letras de Hoje*, *31*(2), 41–50.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bisol, L. (2003). Sandhi in Brazilian Portuguese. *Probus*, *15*(2), 177–200.

Bisol, L. (2013). Sândi vocálico externo. In M. B. Abaurre (Ed.), *Gramática do português culto falado no Brasil: A construção fonológica da palavra vol. VII* (pp. 53–74). Contexto.

Boersma, P., & Weenink, D. (2019). *PRAAT: Doing phonetics by computer* (Version 6.0.53) [Computer software]. http://www.fon.hum.uva.nl/praat/

Collischonn, G. (2001). A sílaba em português. In L. Bisol (Ed.), *Introdução a estudos de fonologia do português brasileiro*. EDIPUCRS.

Faraway, J. J. (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. CRC Press.

Lüdecke, D. (2018). *sjPlot: Data visualization for statistics in social science* (Version 2.8.11) [R package]. https://strengejacke.github.io/sjPlot/

Nespor, M., & Vogel, I. (1986). Prosodic phonology. Foris.

Oliveira Jr., M., & Santos, J. P. M. L. (2018). Análise das vogais átonas finais /e/ e /o/ em sândi vocálico externo em dados do Projeto NURC-Recife. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, *34*(4), 1243–1274.

Tenani, L. E. (2002). *Domínios prosódicos no português do Brasil: Implicações para a prosódia e para a aplicação os processos fonológicos* [PhD dissertation]. Universdade Estadual de Campinas, Brazil.

Tenani, L. E. (2004). O bloqueio do sândi vocálico em PB e PE: Evidências da frase fonológica. *Organon*, *18*(36), 17–29.

Wickham, H., & Seidel, D. (2022). *scales: Scale functions for visualization*. https://scales.r-lib.org

Phonetic accuracy in French learners of English: towards a bilingual database combining articulatory MRI and audio

Alice Léger[1], Coline Caillol[1], Emmanuel Ferragne[1], Hannah King[1], Sylvain Charron[2], Clément Debacker[2,3], Maliesse Lui[2,3], Catherine Oppenheim[2,3]

[1]*Université Paris Cité, CLILLAC-ARP, F-75013 Paris, France*

[2]*Université Paris Cité, Inserm, Institute of Psychiatry and Neurosciences of Paris, F-75014 Paris, France*

[3]*GHU-Paris Psychiatrie et Neurosciences, Hôpital Sainte-Anne, F-75014 Paris, France.*

coline.caillol@etu.u-paris.fr, alice.leger@etu.u-paris.fr

Introduction.

Over the past 15 years, High Temporal Resolution Magnetic Resonance Imaging (HTR-MRI) has emerged as the optimal technique for visualizing all the articulators and structures involved in speech production (Isaieva et al. 2021; Lim et al. 2021). However, it has often been mentioned in the literature that the number of publicly-shared HTR-MRI databases was limited (Isaieva et al. 2021; Belyk, Carignan, and McGettigan 2023). We were able to identify 13 such datasets, and none of them were designed to address bilingualism specifically[1]. This paper therefore describes the early stages of developing such a database of French learners of English with synchronized audio and HTR-MRI. We focus here on the production of three English sounds by advanced L2 English speakers and a native speaker of English – /t/, /r/, and the dark allophone of /l/. Canonical realizations of these sounds in standard varieties of English involve articulatory gestures that are absent from French. Therefore, direct articulatory transfer cannot be used if learners wish to achieve native-like productions. For instance, the place of articulation of [t] has been described as alveolar in English and dental in French, with a potential difference in tongue contact area: apical in English, laminal in French (Dart 1998). Regarding English /r/, no articulatory gesture in French can be transferred to produce the typical alveolar approximant [ɹ]. Its complex coordination of labial, palatal, and pharyngeal constrictions (Harper, Goldstein, and Narayanan 2016) might thus be challenging for learners, especially as native speakers can develop retroflex or bunched tongue shapes to achieve the same low F3 target. As for English dark /l/, [ɬ] has no counterpart in French either. While French /l/ is consistently produced with a single apico-alveolar constriction, the dark /l/ occurs in many standard varieties of English in syllable codas and is realized with a retraction of the back of the tongue followed by an apico-alveolar gesture (Sproat and Fujimura 1993). We provide preliminary descriptions of HTR-MRI data collected from three French advanced speakers of English and a native speaker of English. We investigate whether the L1 French speakers display native-like speech gestures in their L2; that is, whether they developed i) a different place of articulation and tongue contact area for English and French /t/, ii) bunched and/or retroflex tongue shapes for English [ɹ], and iii) a distinct double articulation for the English dark /l/.

Methods.

Participants were three French advanced speakers of English and one native British English speaker. All teach English phonetics at university. The linguistic material consisted of word lists and two small texts, in French and English. MRI data was acquired on a 3T MR scanner (Vantage Galan 3T XGO; Canon Medical Systems, Tochigi, Japan). Mid-sagittal images of the vocal tract were acquired using a single slice 2D fast gradient echo sequence with TR = 2.8 ms, TE = 1.2 ms, BW = 781.25 Hz, FOV = 24×24 cm, flip angle = 5 degrees. In-plane resolution was 2.5×2.5 mm and slice thickness was 10 mm. Images were acquired at a rate of 10 fps with a 16-channel array head-neck coil combined with a flex coil placed over the neck and the mouth. The MRI data was reconstructed with Deep Learning Reconstruction to denoise the images (AiCE by Canon). The audio was captured with a FOMRI III+ microphone by Optoacoustics and further denoised with iZotope.

[1]See shared document here: https://tinyurl.com/MRIDatabaseReview

Results and discussion.

Regarding /t/, no difference in place of articulation was observed: all participants produced an alveolar constriction in both English and French. Three of them, native speaker included, produced an apical consonant in both languages. The other, however, showed laminal contact in French and an apico-laminal intermediate contact in English, suggesting a potential influence of the L1 on the L2. Overall results also imply that the lamino-dental (French /t/) versus apico-alveolar (English /t/) distinction may not be as clear-cut as traditionally thought, and could be the result of inter-speaker variability within and across both languages. Turning to /r/, we found a native-like diversity of lingual patterns consistent with prior findings (Léger, King, and Ferragne 2023). One speaker exclusively used retroflex shapes, another used bunched shapes, and the third one used both; In addition, L2 English speakers demonstrated coarticulation patterns similar to that of native speakers (Mielke, Baker, and Archangeli 2016). Finally, all participants displayed a dark /l/ in English, with two constrictions at the tongue dorsum and tip in coda /l/ (Fig 1.5 to 1.8). Native and non-native speakers of British English produced a clear onset /l/ in French and English, characterized by a single alveolar constriction at the tongue tip (Fig 1.1, 1.2, 1.4). The one L2 speaker with an American accent, however, velarised /l/ in that position (Fig 1.3), in line with the darker onset productions observed in native speakers of that variety (Recasens 2012). Interestingly, the characteristic "saddle" shape of the dark /l/ (Wrench and Scobbie 2003) was visually less salient in one L2 English speaker, who displayed an atypical constriction at the pharynx rather than the velum (Fig 1.6), indicating an alternative articulatory strategy for darkness. These preliminary observations will be discussed more thoroughly in the final presentation.

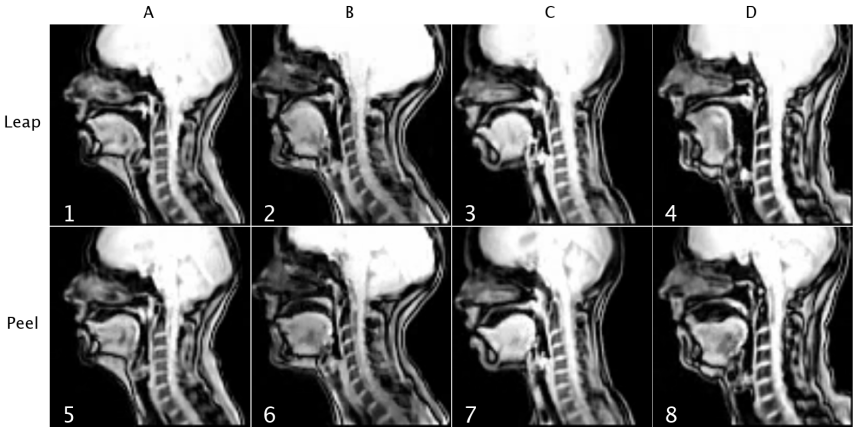


Figure 1: /l/ in onset and coda position for the native speaker (A) and English L2 speakers (B, C, D).

References.

Belyk, Michel, Christopher Carignan, and Carolyn McGettigan (July 2023). "An Open-Source Toolbox for Measuring Vocal Tract Shape from Real-Time Magnetic Resonance Images". In: *Behavior Research Methods*. DOI: 10.3758/s13428-023-02171-9.

Dart, Sarah N. (Jan. 1998). "Comparing French and English Coronal Consonant Articulation". In: *Journal of Phonetics* 26.1, pp. 71–94. DOI: 10.1006/jpho.1997.0060.

Harper, Sarah, Louis Goldstein, and Shrikanth S. Narayanan (Sept. 2016). "L2 Acquisition and Production of the English Rhotic Pharyngeal Gesture". In: *Interspeech 2016*. ISCA, pp. 208–212. DOI: 10.21437/Interspeech.2016-658.

Isaieva, Karyna, Yves Laprie, Justine Leclère, Ioannis K. Douros, Jacques Felblinger, and Pierre-André Vuissoz (Oct. 2021). "Multimodal Dataset of Real-Time 2D and Static 3D MRI of Healthy French Speakers". In: *Scientific Data* 8.1, p. 258. DOI: 10.1038/s41597-021-01041-3.

Léger, Alice, Hannah King, and Emmanuel Ferragne (2023). "Is Rhoticity On The Tip Of Your Tongue? Investigating Tongue Shapes For English /r/ In French Learners With Ultrasound". In: *ICPhS*. Prague, pp. 2741–2745.

Lim, Yongwan et al. (July 2021). "A Multispeaker Dataset of Raw and Reconstructed Speech Production Real-Time MRI Video and 3D Volumetric Images". In: *Scientific Data* 8.1, p. 187. DOI: 10.1038/s41597-021-00976-x.

Mielke, Jeff, Adam Baker, and Diana Archangeli (2016). "Individual-Level Contact Limits Phonological Complexity: Evidence from Bunched and Retroflex /r/". In: *Language*, pp. 101–140. DOI: 10.1353/lan.2016.0019.

Recasens, Daniel (2012). "A Cross-Language Acoustic Study of Initial and Final Allophones of /l/". In: *Speech Communication* 54.3, pp. 368–383. DOI: 10.1016/j.specom.2011.10.001.

Sproat, Richard and Osamu Fujimura (July 1993). "Allophonic Variation in English /l/ and Its Implications for Phonetic Implementation". In: *Journal of Phonetics* 21.3, pp. 291–311. DOI: 10.1016/S0095-4470(19)31340-3.

Wrench, Alan A and James M Scobbie (2003). "Categorising Vocalisation of English /l/ using EPG, EMA and Ultrasound." In: *Proceedings of the 6th international Seminar on Speech Production*.

Spatio-temporal properties of Japanese coronal consonants: An ultrasound study of /d/ and /r/

Maho Morimoto[1], Takayuki Nagamine[2]

[1]Sophia University/Japan Society for the Promotion of Science
[2]Lancaster University
maho.morimoto.jp@gmail.com, t.nagamine@lancaster.ac.uk

Introduction. Previous literature on Japanese consonants has noted the similarity of the liquid consonant /r/ and the alveolar plosive consonant /d/. While researchers generally agree that there are several variants of /r/ including one like a 'weak [d]' (e.g., Kawakami 1977), some argue that /r/ is articulatorily different from /d/ in that /r/ involves a ballistic gesture (e.g., Akamatsu 1997). It has also been pointed out that there is a varying degree of similarity between these two consonants depending on the context, with more similarity exhibited in phrase-initial and post-nasal environments (e.g., Arai 2013). While several studies using electropalatograhy (EPG) have demonstrated that /r/ is not just a short version of /d/ (e.g., Kawahara et al. 2017; Kochetov 2017), the exact articulatory mechanisms underlying the similarities and differences between /r/ and /d/ are not well understood, especially of the tongue dorsum movements that are not well-captured using EPG. In this study, we investigate articulatory differences between /d/ and /r/ in Japanese. We use ultrasound to obtain clear images of tongue dorsum, whose behavior might differ depending on the vocalic context.

Methods. We report results from one 21-year-old male speaker from Tokyo. The participant produced the following Japanese words containing intervocalic /d/ and /r/: /ada/ ("avenge"), /ara/ ("coarseness"), /badi:/ ("body/buddy"), /bari:/ ("Barry"), /kaNdou/ ("sensation"), /kaNro/ ("honeydew"). They were repeated five times in random order, resulting in a total of 30 tokens of /d/ and /r/ for analysis. We obtained audio recordings (at 22,050 Hz) and midsagittal ultrasound tongue images (at approximately 113 fps) using Articulate Assistant Advanced (AAA) version 221.0.0 (Articulate Instruments 2023). The probe was stabilized using a headset (Spreafico et al. 2018). Prior to analysis, we automatically segmented /d/ and /r/ using Montreal Forced Aligner (McAuliffe et al. 2017) and then manually adjusted the boundaries wherever necessary using Praat (Boersma and Weenink 2022). Tongue splines were automatically fitted using the DeepLabCut (DLC) plug-in on AAA based on the acoustic consonantal intervals. DLC estimates tongue splines based on 11 x/y coordinates in each ultrasound frame. The tongue contour data was extracted at 11 equidistant time points during the target consonants /d/ and /r/. The tongue splines were rotated and offset using the speaker's occlusal plane that we measured by having the speaker bite a thin plastic plate (Scobbie et al. 2011). To identify primary variation in midsagittal tongue movement in /d/ and /r/, we ran principal components analysis (PCA) using scripts publicly available from Nance and Kirkham (2021). PCA was run based on the *z*-scored x/y coordinates from all tongue splines extracted for /d/ and /r/, and we tracked the time-varying changes of the first two PCs that accounted for the largest proportion of variance to visually inspect how tongue movement differs between /d/ and /r/.

Results. The left panel in **Figure 1** shows time-varying changes in midsagittal tongue shapes for /d/ (top) and /r/ (bottom). The results of PCA are shown in the right panel in **Figure 1**. Variations explained by PCs 1 and 2 (top right) are superimposed on the tongue midsagittal tongue shape, in which the mean tongue shape is represented with the bold line and the variation captured by each PC with dashed (plus) and dotted (minus) lines by adding and subtracting a standard deviation associated with each PC from the mean tongue curve. Shown in bottom right in **Figure 1** are time-varying changes in each PC dimension during each consonant. The consonant duration is normalized and expressed proportionally between 0% (consonantal onset) and 100% (consonantal offset). Thin lines represent PC changes of each token, with the thick lines smoothing them and the dotted lines showing the 95% confidence interval.

The midsagittal tongue shape in the left panel in **Figure 1** suggests that there are some differences in the dynamic spatial properties of /d/ and /r/. We have found that the variation in the tongue motion of the two consonants can be described in terms of two principal components, PC1 (76.85%) and PC2 (10.97%). In **Figure 1**, PC1 appears to capture the tongue retraction component at the tongue dorsum, correlated with the height of the tongue body. The time-varying changes for PC1 (bottom right) show that the tongue dorsum for /r/ maintains a retracted tongue position when flanked by low vowels, while /d/ transitions from an anterior tongue dorsum position to one comparable to /r/ at the offset. In intervocalic position /a_i/, we observe that the PC1 changes were relatively small for /r/ compared to /d/, which might suggest a dorsal stabilization mechanism for /r/. The PC1 changes for /d/ and /r/ in /aN_o/ context are largely comparable with the two trajectories overlapping for the majority of consonantal intervals. PC2, on the other hand, suggests a very slight variation around the tongue body, which is slightly raised for /r/ across vocalic contexts. The difference in PC2 between /d/ and /r/ spans throughout the consonantal intervals.

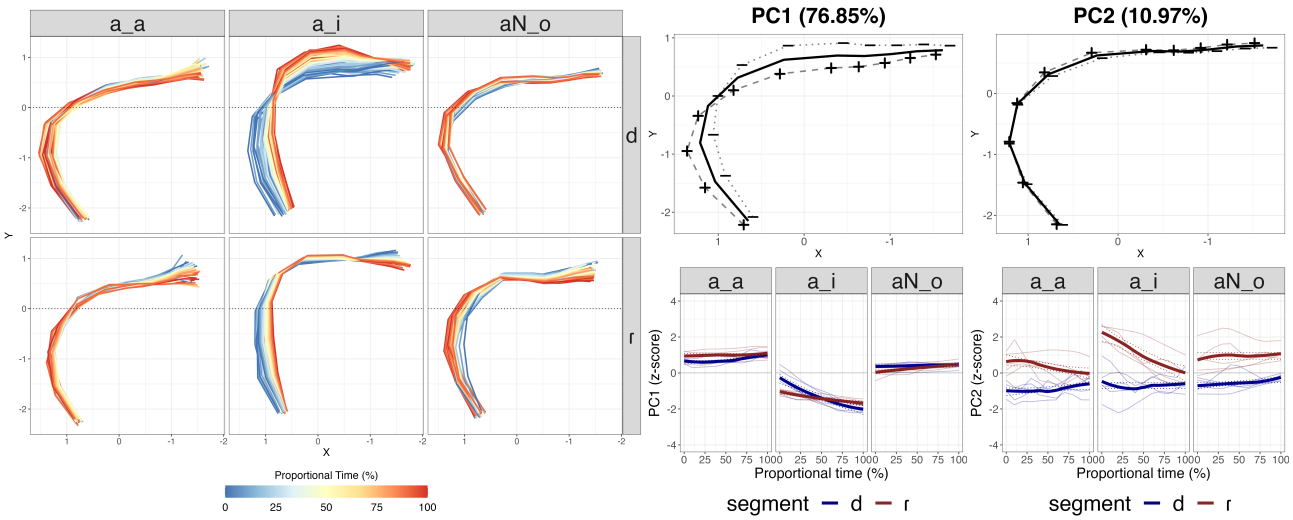


Figure 1: *Left: midsagittal tongue shapes during the consonant intervals in each vowel context for /d/ and /r/. Tongue tip points to the right; Right: variation captured in PCs 1 and 2 (top) and time-varying changes of each PC (bottom).*

Discussion. The current study highlights the possible differences in the articulation of Japanese /d/ and /r/. First, we suggest that one of the key articulatory differences between /d/ and /r/ lies in tongue retraction and stabilization. The tongue retraction in /r/ is evident in the overall posterior tongue dorsum in /a_a/ context. In addition, as seen in the midsagittal tongue shape and the dynamic changes in PC1 in **Figure 1**, the relative stability in the tongue dorsum position for /r/ in /a_i/ context points to some dorsal stabilization mechanism of /r/, while /d/ is more susceptible to vowel coarticulation. The similarity in the degree of tongue retraction in /d/ and /r/ in /aN_o/ context is consistent with previous literature pointing out that the two consonants are especially confusable after coda nasals. It is noteworthy that the duration of the two consonants was also largely comparable in this environment, while /d/ was generally longer than /r/ in other contexts. Finally, the slight raising of the tongue body in /r/ as suggested by PC2 may be a by-product of tongue body compression as a result of tip retraction in /r/, which might result from different manner requirements for /d/ and /r/.

Although it only considers a small number of tokens, the current study demonstrates that ultrasound paired with PCA allows us to better investigate articulatory mechanisms of Japanese coronal consonants. Future research will incorporate a larger number of speakers as the current study is based on the productions of a single speaker. In addition, it is necessary to examine the productions of /d/ and /r/ in a wider variety of contexts, as articulation of /d/ and /r/ may be influenced by prosodic positions and adjacent vowels (Yamane et al. 2015; Maekawa 2023). Note also that the current methodology may not fully account for tongue tip movement or jaw displacement. These considerations will help to better characterize the articulation of Japanese coronal consonants.

References

Akamatsu, T. (1997). “Japanese phonetics: Theory and practice (Vol. 3)”. Lincom Europa.

Arai, T. (2013). “On why Japanese /r/ sounds are difficult for children to acquire”. In: *Interspeech 2013*. Lyon, France: International Speech Communication Association, pp. 2445–2449.

Articulate Instruments. (2023). “Articulate Assistant Advanced version 221.0.0”. [Computer software]. Edinburgh, Articulate Instruments.

Boersma, P., & Weenink, D. (2022). “Praat: Doing phonetics by computer version 6.2.19”. [Computer software].

Kawahara, S., Matsui, M., & Shaw, J. (2017). “Some aspects of Japanese consonant articulation: A preliminary EPG study”. In: *ICU Working Papers in Linguistics II*, pp. 1–12.

Kawakami, S. (1977). “Nihongo onsei gaisetsu [Outline of Japanese phonetics]”. Tokyo: Ōfūsha.

Kochetov, A. (2017). “Linguopalatal contact contrasts in the production of Japanese consonants: Electropalatographic data from five speakers”. In: *Acoustical Science and Technology* 39.2, pp. 84–91.

Maekawa, K. (2023). “Articulatory characteristics of the Japanese /r/: A real-time MRI study”. In: *Proceedings of the 20th International Congress of Phonetic Sciences*. Prague: Guarant International, pp. 992–996.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi”. In: *Interspeech 2017*. Stockholm, Sweden: International Speech Communication Association, pp. 498–502.

Nance, C., & Kirkham, S. (2022). “Phonetic typology and articulatory constraints: The realization of secondary articulations in Scottish Gaelic rhotics”. In: *Language* 98.3, pp. 419–460.

Scobbie, J., Lawson, E., Cowen, S., Cleland, J., & Wrench, A. (2011). “A common co-ordinate system for mid-sagittal articulatory measurement”. In: *QMU CASL Working Papers* 20, pp. 1–4.

Spreafico, L., Pucher, M., & Matosova, A. (2018). “UltraFit: A speaker-friendly headset for ultrasound recordings in speech science.” In: *Interspeech 2018*. Hyderabad, India: International Speech Communication Association, pp. 1517–1520.

Yamane, N., Howson, P., & Wei, P.-C. G. (2015). “An ultrasound examination of taps in Japanese”. In: *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: The International Phonetic Association, pp. 1–5.

Comparing the real-time perception of French nasal and labial coarticulation

Francesco Rodriquez[1], Marianne Pouplier[1], Phil J. Howson[1], Eva Reinisch[2], Justin J.H. Lo[3], Christopher Carignan[4], Bronwen G. Evans[4]

[1]Ludwig-Maximilians-Universität München, Germany
[2]Austrian Academy of Sciences, Vienna, Austria
[3]Lancaster University, United Kingdom
[4]University College London, United Kingdom

{f.rodriquez|pouplier|p.howson}@phonetik.uni-muenchen.de,
Eva.Reinisch@oeaw.ac.at, j.h.lo@lancaster.ac.uk,
{c.carignan|bronwen.evans}@ucl.ac.uk

Introduction. An integral feature of speech production is the overlap of articulatory gestures in time, a phenomenon also known as coarticulation. Coarticulatory information permeates the acoustic signal and is actively sought out by listeners in spoken-word recognition (Fowler & Brown 2000). In French, it is reported that fine-grained differences in coarticulatory timing for both nasality and lip rounding provide perceivable information to listeners for phoneme identification and disambiguation (Desmeules-Trudel & Zamuner 2019; Benguerel & Adelman 1976; Hirsch et al. 2003). An intuitive explanation would be to attribute the sensitivity to these coarticulatory cues to the contrastive nature of nasality and lip rounding in the French vowel system. However, anticipatory lip rounding and nasal coarticulation are known to differ in their temporal extent in French: while anticipatory lip rounding was found to be quite extensive and variable (e.g., Noiray et al. 2010) the temporal extent of nasal coarticulation is rather constrained (e.g., Pouplier et al. 2023). This difference in the temporal extent of anticipatory coarticulation within the same language affords the unique opportunity to investigate whether these inter-articulator production differences are directly reflected in listeners' perceptual patterns. Also, the question of whether information from coarticulation with the same phonological status (here: contrastive) can be expected to be treated in the same way by listeners has not been directly investigated yet. The results of such an inquiry could shed light on the complex dynamics of the production-perception link in coarticulation (e.g., Beddor et al. 2018).

The present study investigates the use of coarticulatory timing cues in French by examining listeners' perceptual sensitivity to temporal variation in anticipatory nasalization and anticipatory lip rounding. We ask whether listeners are sensitive to fine-grained temporal differences in coarticulatory timing during spoken-word recognition as soon as the information becomes available in the signal. Listeners' responses should then systematically vary with coarticulatory onset for both articulators.

Methods. 16 native speakers/listeners of French participated in both a production (reading task) and a perception experiment (visual-world eye-tracking experiment). *Production experiment:* Nasal coarticulation data was recorded with a nasalance device and nasality was measured as (mean-normalized) nasal channel intensity. Anticipatory lip rounding data was quantified off video recordings (Lallouache, 1991). Rounding was measured as the distance between the lip corners (lip spread). The target stimuli, embedded in a carrier phrase, contain minimal pairs distinguished by a nasal/oral consonant in VN/VC sequences (e.g., *l'aîné* [le.ne] vs. *l'été* [le.te]) or by a rounded/unrounded vowel (e.g., *Caire* [kɛʁ] vs. *cœur* [kœʁ]) of the following front vowel pairs: /e/~/ø/, /ɛ/~/œ/, /i/~/y/. The onset of coarticulation was determined algorithmically (as explained in Lo et al. 2023) as a point in signal divergence between a given nasal/oral (rounded/unrounded) minimal pair.

Perception experiment: Tokens for the eye-tracking experiment were selected per articulator by sampling the extremes ($1^{st}/4^{th}$ quartile) of the coarticulation onset point distributions across speakers. For each articulator the same number of tokens with extensive and constrained coarticulation was selected. The extensive category's mean coarticulation onset preceded that target segment by 288ms for lip rounding and by 154ms for nasality. The constrained category's mean coarticulation onset preceded the target segment by 110ms for lip rounding and 47ms for nasality. Each target was presented in its original carrier phrase (114 unique trials), each presented twice and mixed with about the same number of fillers. Visual referents were printed words of the minimal pairs with targets presented once on the left and once on the right. Participants clicked on the word that they heard (two-alternative forced choice task). The probability of target fixations over time (5ms bins) was calculated by means of Growth Curve Analysis (GCA) (Mirman et al. 2008). One model per articulator was computed including a fixed effect of coarticulation type (extensive, constrained), functions for time ($time^n$, n = [1,7]) as well as the interactions between the two. Smoothing splines were used to obtain 0.95 confidence intervals for fixation proportions over time as well as to determine differences in the growth curves for the extensive and constrained condition (Wendt et al. 2014).

Results. The growth curves of the proportions of fixation on the correct target word are shown in **Figure 1** (A-B). The extensive and constrained conditions for an articulator are compared in divergence plots (**Figure 1**, C-D), where

significant divergences in correct target fixations between the extensive and constrained condition are marked in red. For nasality, the target fixation curve in the extensive condition rises ~50ms after target segment onset. Considering 200ms needed to perform a saccade (Travis 1936), the results suggest that listeners use extensive nasalization cues as soon as they are available in the signal (mean anticipatory interval of nasality in extensive condition: 154ms) (**Figure 1**, A). Responses in the constrained condition align with the extensive condition, so the curves do not significantly diverge (**Figure 1**, C). This is surprising as it greatly exceeds the time window at which nasalization cues become available in the constrained condition (anticipatory interval: 18ms to 59ms). Possibly, this is due to confounds or unresolved biases currently under investigation. For lip rounding, target fixations in the extensive condition start rising at target onset (= zero point), suggesting a listener response slightly delayed with respect to the mean onset of coarticulation (mean anticipatory interval of rounding in extensive condition: 288ms) (**Figure 1**, B). The constrained condition significantly diverges from the extensive condition at 50ms into the target (**Figure 1**, D).

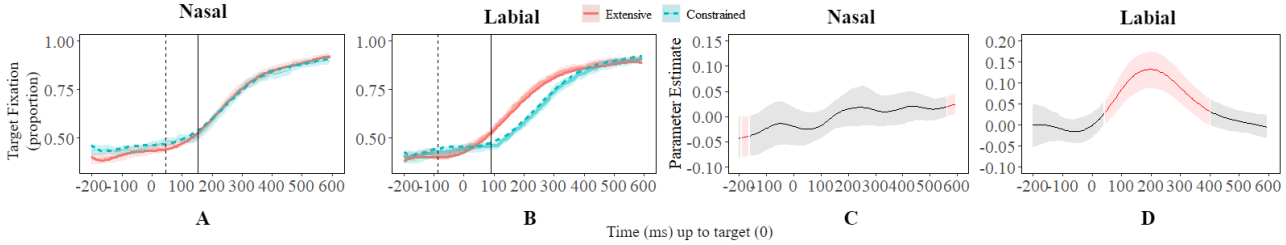


Figure 1: *Growth curve analysis for extensive (red solid line) and constrained (blue dotted line) tokens for nasal (A) and labial (B) perception. Mean fixation proportions in opaque colors. Estimated earliest point of reaction in perception (mean coarticulatory onset up to target across produced stimuli + 200ms for saccade execution), for the extensive (vertical dashed line, A-B) and constrained condition (vertical solid line, A-B). Divergence in proportion of fixation between extensive and constrained coarticulation for nasal (C) and labial (D) coarticulation. Significant differences are marked in red. Target onset at 0.*

Discussion. The results suggest that French listeners are sensitive to extensive coarticulatory cues as soon as they become available in the acoustic signal for both nasal and labial coarticulation. The time course of perception thus mirrors the fine-grained differences in the temporal extent of coarticulatory information in production. This observation might support the idea that French listeners are attuned to anticipatory nasalization and lip rounding as real-time cues because of their phonologically contrastive status. Notably, in French this seems to hold despite the observation that anticipatory lip rounding is more variable (i.e., the cue is less consistent) than anticipatory nasalization. While we focused on group-level perceptual patterns here, in planned analyses we intend to compare speaker-specific production-perception patterns (as in Beddor et al. 2018) to test whether an individual speakers' coarticulatory behavior in production (e.g., early vs. late 'coarticulators') is reflective of their use of coarticulatory information in perception (e.g., early vs. late use of coarticulatory cues).

References

Beddor, P., Coetzee, A., Styler, W., McGowan, K., & Boland, J. (2018) The time course of individuals' perception of coarticulatory information is linked to their production: implications for sound change. Language, 94(4), 931-968.

Benguerel, A.P., & Adelman, S. (1976). Perception of coarticulated lip rounding. Phonetica, 33(2), 113-126.

Desmeules-Trudel, F., & Zamuner, T.S. (2019). Gradient and categorical patterns of spoken-word recognition and processing of phonetic details. Attention, Perception, & Psychophysics, 81(5), 1654-1672.

Fowler, C.A., & Brown, J.M. (2000). Perceptual parsing of acoustic consequences of velum lowering from information for vowels. Perception & Psychophysics, 62, 21-32.

Hirsch, F., Sock, R., Connan, P., & Brock, G. (2003). Auditory effects of anticipatory rounding in relation with vowel height in French. Proceedings of the 15th International Congress of Phonetic Sciences, 1445-1448.

Lallouache, M. T. (1991). Un poste visage-parole couleur: Acquisition et traitement automatique des contours des lèvres. Ph.D. dissertation, Institut National Polytechnique de Grenoble.

Lo, J.J.H., Carignan, C., Pouplier, M., Alderton, R., Rodriquez, F., Evans, B.G., Reinisch, E. 2023. Language specificity vs. speaker variability of anticipatory labial coarticulation in German and English. *Proceedings of the 20th International Congress of Phonetic Sciences,* 2105-2109.

Travis, R. C. (1936). The latency and velocity of the eye in saccadic movements. *Psychological Monographs*, 242-249.

Mirman, D. Dixon, J.A., & Magnuson, J.S. (2008). Statistical and computational models of the visual world paradigm: growth curves and individual differences. *Journal of Memory and Language, 59*(4), 475-494.

Pouplier, M., Rodriquez, F., Alderton, R., Lo, J.J.H., Reinisch, E., Evans, B.G., Carignan, C. (2023). The window of opportunity: Anticipatory nasal coarticulation in three languages. *Proceedings of the 20th International Congress of Phonetic Sciences* (Prague, Czech Republic), 2085-2089.

Noiray, A., Cathiard, M., Abry, C., & Ménard, L. (2010). Lip rounding anticipatory control: Crosslinguistically lawful and ontogenetically attuned. In Maassen, B. & van Lieshout, P. (eds.), *Speech motor control: New developments in basic and applied research*, Oxford, United Kingdom: Oxford University Press, 153-171.

Wendt, D., Brand, T., & Kollmeier, B. (2014). An Eye-Tracking Paradigm for Analyzing the Processing Time of Sentences with Different Linguistic Complexities. *PLoS ONE*, e100186

Relationship between working memory and auditory rhythm discrimination in adults who stutter

Emily O. Garnett[1], Bailey Rann[2], Nicholas Mularoni[1], Toni Smith[2], Soo-Eun Chang[1], J. Devin McAuley[2]

[1]University of Michigan
[2]Michigan State University

emilyog@med.umich.edu, rannbail@msu.edu, nmula@med.umich.edu, smit3094@msu.edu, sooeunc@med.umich.edu, dmcauley@msu.edu

Introduction. Children and adults who stutter show poorer auditory rhythm discrimination than adults who do not stutter, especially for complex rhythms that don't have a consistently marked beat (Wieland et al., 2015; Garnett et al., 2023). This suggests that stuttering may involve a deficit in internal generation of a periodic beat, known as the internal beat-deficit hypothesis (Alm, 2004; Garnett et al., 2023). Entrainment models of short-interval assume that rhythm discrimination leverages an oscillatory mechanism that is entrained by the beat of the to-be-discriminated rhythms whereas interval models of timing assume that rhythm discrimination is based on an interval-by-interval comparison of the rhythms (McAuley & Jones, 2003). If individuals who stutter rely to a greater degree on interval-by-interval duration comparisons to discriminate rhythms than beat-based timing, a stronger relationship between working memory and rhythm discrimination performance might be expected for adults who stutter (AWS) compared to adults who do not stutter. To test this hypothesis, the current study examined the relationship between working memory and rhythm discrimination in AWS and adults who do not stutter (controls). Data was combined from three different datasets where participants performed the same rhythm discrimination and working memory tasks across all datasets. The relationship between working memory and rhythm discrimination was then examined in AWS and controls.

Methods. Participants included 64 AWS (mean age: 28, 26F) and 91 Controls (Mean age: 27, 39F). Working memory was measured using an automated Operation Span (OSPAN) Task (Unsworth et al., 2005) during which participants solved a series of arithmetic equations while remembering lists of unrelated letters. Participants were presented with one equation at a time and asked to verify whether the equation is correct. Between equations, participants were shown a letter and at the end of the sequence they selected the letters in the order that they were shown. In the present study, OSPAN scores provide a measure of working memory. In the Rhythm Discrimination Task, participants heard two successive presentations of a standard rhythm and judged whether a third comparison rhythm was the same or different from the standard rhythm. The rhythms were either simple or complex. Simple and complex rhythms were comprised of the same number of intervals but were arranged in different ways. We used similar simple and complex rhythms to prior studies, where simple rhythms had tones marking every beat whereas complex rhythms lacked a regular beat. Rhythm discrimination performance was assessed using the signal detection measure d'.

Results. OSPAN scores did not differ between groups (AWS: M=52, SD=15; Control: M=56, SD=15; p=0.12). There was also no significant difference in rhythm discrimination between groups for either simple rhythms (AWS: M=2.31, SD=1.07; Control: M=2.29, SD=0.97; p=0.9) or complex rhythms (AWS: M=1.58, SD=1.03; Control: M=1.85, SD=0.89; p=0.10). Figure 1 depicts the relationship between working memory and rhythm discrimination for AWS and controls. In the left panel showing simple rhythm discrimination, there is a significant positive correlation between working memory and simple rhythm discrimination for the AWS group but not the Control group. Using Fischer r-to-z transformation, the group difference between correlations did not reach statistical significance ($z=-1.12$, $p=0.13$). In the right panel showing complex rhythm discrimination, there is a significant correlation between working memory and complex rhythm discrimination for AWS only, and the relationship is significantly greater compared to controls ($z=-1.7$, $p<0.04$).

Discussion. As hypothesized, AWS showed a stronger relationship between working memory and rhythm discrimination compared to adults who do not stutter. This relationship was significantly greater in AWS compared to controls for complex rhythms. The findings cannot be explained by group differences in working memory or rhythm discrimination, though AWS scored marginally lower in both simple and complex rhythm discrimination. As predicted by the internal beat-deficit hypothesis and consistent with entrainment models of short-interval timing, AWS do not appear to engage beat-based timing mechanisms to the same degree as adults who do not stutter. Rather, our findings support the use of an interval-by-interval timing mechanism for AWS, even when discriminating simple rhythms in which tones fall on a periodic beat, suggesting that AWS rely on working memory to a greater degree than control participants to discriminate rhythms. Beat- and interval-based timing are supported by different neural architecture: the basal ganglia thalamocortical (BGTC) network (e.g., supplementary motor area, putamen) supports beat-based timing, whereas the cerebellum supports

interval-based timing (Breska & Ivry, 2018; Grahn & Brett, 2009; Grahn & McAuley, 2009; Grube et al., 2010; Nozaradan et al., 2017; Teki et al., 2011). Furthermore, these proposed beat-based and interval-based timing networks overlap with internal and external timing mechanisms, respectively. It is well established that the use of external pacing (e.g., a metronome) during speech markedly decreases stuttering, likely related to a reduced reliance on disrupted "internal timing." Structural and functional abnormalities have been found in the BGTC network in stuttering (e.g., Chang & Guenther, 2019), which thus may be reflected in both speech and nonspeech related deficits. Based on these results, we conclude that AWS can discriminate rhythms at close to the same level as adults who do not stutter, but by relying to a greater degree on interval-based timing mechanism that leverages working memory.

.

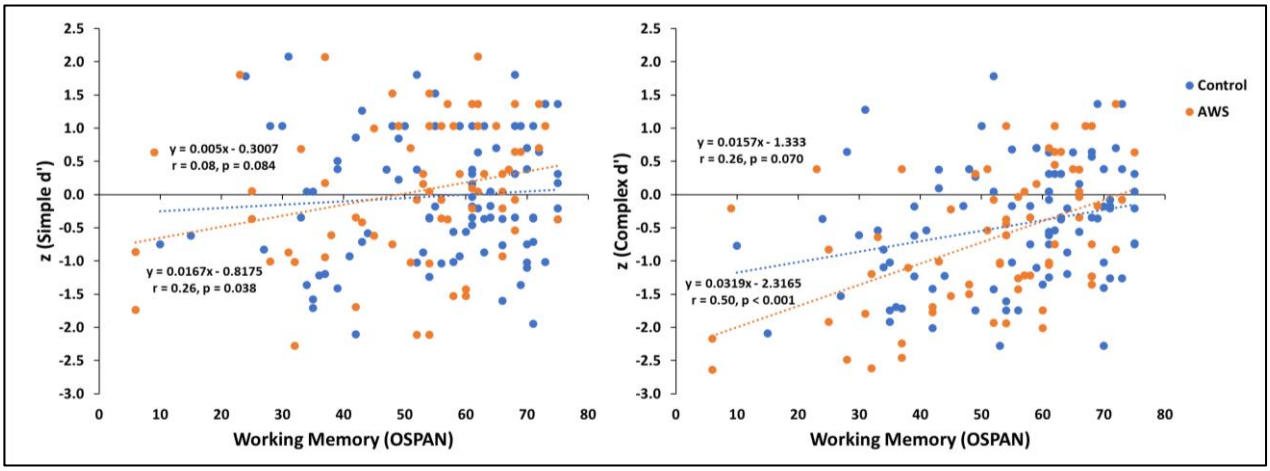


Figure 1: *AWS show a significant positive correlation between rhythm discrimination (d') and working memory for both simple (left) and complex (right) rhythms.*

References

Alm, P. A. (2004). Stuttering and the basal ganglia circuits: a critical review of possible relations. Journal of communication disorders, 37(4), 325-369.

Breska, A., & Ivry, R. B. (2018). Double dissociation of single-interval and rhythmic temporal prediction in cerebellar degeneration and Parkinson's disease. Proceedings of the National Academy of Sciences, 115(48), 12283-12288.

Chang, S. E., & Guenther, F. H. (2020). Involvement of the cortico-basal ganglia-thalamocortical loop in developmental stuttering. Frontiers in Psychology, 10, 3088.

Garnett, E. O., McAuley, J. D., Wieland, E. A., Chow, H. M., Zhu, D. C., Dilley, L. C., & Chang, S.-E. (2023). Auditory rhythm discrimination in adults who stutter: An fMRI study. Brain and Language, 236, 105219. https://doi.org/10.1016/j.bandl.2022.105219

Grahn, J. A., & Brett, M. (2007). Rhythm and beat perception in motor areas of the brain. Journal of cognitive neuroscience, 19(5), 893-906.

Grahn, J. A., & Brett, M. (2009). Impairment of beat-based rhythm discrimination in Parkinson's disease. cortex, 45(1), 54-61.

Grube, M., Cooper, F. E., Chinnery, P. F., & Griffiths, T. D. (2010). Dissociation of duration-based and beat-based auditory timing in cerebellar degeneration. Proceedings of the National Academy of Sciences, 107(25), 11597-11601.

McAuley, J. D., & Jones, M. R. (2003). Modeling effects of rhythmic context on perceived duration: A comparison of interval and entrainment approaches to short-interval timing. Journal of Experimental Psychology: Human Perception and Performance, 29(6), 1102–1125.

Nozaradan, S., Schwartze, M., Obermeier, C., & Kotz, S. A. (2017). Specific contributions of basal ganglia and cerebellum to the neural tracking of rhythm. Cortex, 95, 156-168.

Povel, D. J., & Essens, P. (1985). Perception of temporal patterns. Music perception, 2(4), 411-440.

Teki, S., Grube, M., Kumar, S., & Griffiths, T. D. (2011). Distinct neural substrates of duration-based and beat-based auditory timing. Journal of Neuroscience, 31(10), 3805-3812.

Unsworth, N., Heitz, R.P., Schrock, J.C. et al. An automated version of the operation span task. Behavior Research Methods 37, 498–505 (2005).

Wieland, E. A., McAuley, J. D., Dilley, L. C., & Chang, S.-E. (2015). Evidence for a rhythm perception deficit in children who stutter. Brain and Language, 144, 26–34. https://doi.org/10.1016/j.bandl.2015.03.008

Identifying different types of lingual tremor in individuals with Parkinson's disease using electromagnetic articulography: a follow-up study

Teja Rebernik[1,2], Jidde Jacobi[1], Mark Tiede[3], Martijn Wieling[1]

[1]*Rijksuniversiteit Groningen*
[2]*Vrije Universiteit Brussel*
[3]*Yale University*
t.rebernik@rug.nl, j.jacobi@rug.nl, mark.tiede@yale.edu, m.b.wieling@rug.nl

Introduction. One of the cardinal symptoms of Parkinson's disease is tremor, which has been defined as the involuntary and oscillatory movement of a body part, and can be classified as either rest tremor (with a frequency of 3-6 Hz) or action tremor (Bhatia et al. 2018; Chan et al. 2022). The latter can be subdivided into postural tremor, occurring at a frequency of 4-9 Hz when a specific posture is actively maintained, and kinetic tremor, occurring at a frequency of 7-12 Hz when voluntary movement is carried out (ibid.). While tremor in PD is most often evaluated in the limbs, some prior studies have assessed resting tremor of the tongue as well. A study by Hunkers and Abbs (1990), using EMG electrodes and displacement transduction devices, showed resting lingual tremor in three individuals with Parkinson's disease (IwPD) with a frequency between 4.5-5 Hz (RMS movement amplitude: 0.2-1.9 mm). Likewise, a study by Jacobi et al. (2020), using NDI-WAVE electromagnetic articulography sensors, reported lingual tremor in two IwPD with a mean measured frequency of 3.7 Hz in the tongue back sensor (RMS movement amplitude: 1.5-1.7 mm) and between 3.7-3.9 Hz in the tongue tip sensor (RMS movement amplitude: 0.2-0.9 mm). We build upon the study by Jacobi et al. (2020) by identifying potential tremor in a wider range of tasks that may elicit different kinds of tremor. To the best of our knowledge, no study has yet assessed lingual action tremor in IwPD.

Methods. We assessed tremor in 33 IwPD who took part in a larger study (approved by our institutional Medical Ethics Review Board). Prior to the experiments taking place, the participants underwent an MDS-UPDRS assessment of motor- and non-motor symptom severity (parts I-III; Goetz et al. 2008). Based on the MDS-UPDRS assessment and following Stebbins et al. (2013), IwPD were assigned to a tremor-dominant (TD), Postural Instability Gait Difficulty (PIGD) or indeterminate phenotype group. All IwPD participated while being *ON* levodopa.

Participants completed several experimental tasks that focused on articulatory kinematics, including a motor learning task with real-time visual feedback, a formant perturbation speech production task, and a sustained vowel phonation task. NDI-VOX EMA sensors were placed following procedures outlined in Rebernik, Jacobi, Jonkers, et al. (2021). Participant data was visually examined in MVIEW during post-processing. The Results section reports on two IwPD who showed pronounced lingual tremor in most recordings for at least one experimental task. Lingual tremor characteristics were assessed using a continuous wavelet transform (CWT; Lilly 2017) applied to the first principal component of the Tongue Tip (TT) sensor trace using the MATLAB Wavelet toolbox. The absolute CWT value was displayed as a function of both time and frequency in a scalogram, allowing us to identify whether a frequency potentially associated with tremor was present in the signal, and where this frequency occurred. For comparability with the prior study by Jacobi et al. (2020), frequency peaks and RMS amplitudes were calculated using the vertical TT sensor trace that had been transformed using a Fast Fourier Transform (FFT). In line with this study and our study on the accuracy of the NDI-VOX device (Rebernik, Jacobi, Tiede, et al. 2021), an RMS amplitude of 0.2 mm was set as the threshold for distinguishing between trials with tremor, and noise.

Example 1. Participant A was a 67-year old male, diagnosed with idiopathic PD five years prior to being included in the study (MDS-UPDRSIII: 30 points; TD phenotype). In the real-time visual feedback task, he showed kinetic tremor (see Figure 1, left) of the tongue tip at 11.5 Hz (SD = 0.1) with an RMS amplitude of 0.5 mm (SD = 0.2). During the speech production and sustained phonation tasks, he showed both resting tremor, with a mean frequency of 3.6 Hz (SD = 0.5) and RMS amplitude of 0.4 mm (SD = 0.09), and kinetic tremor, with a mean frequency of 11.3 Hz (SD = 0.5) and RMS amplitude of 0.3 mm (SD = 0.06).

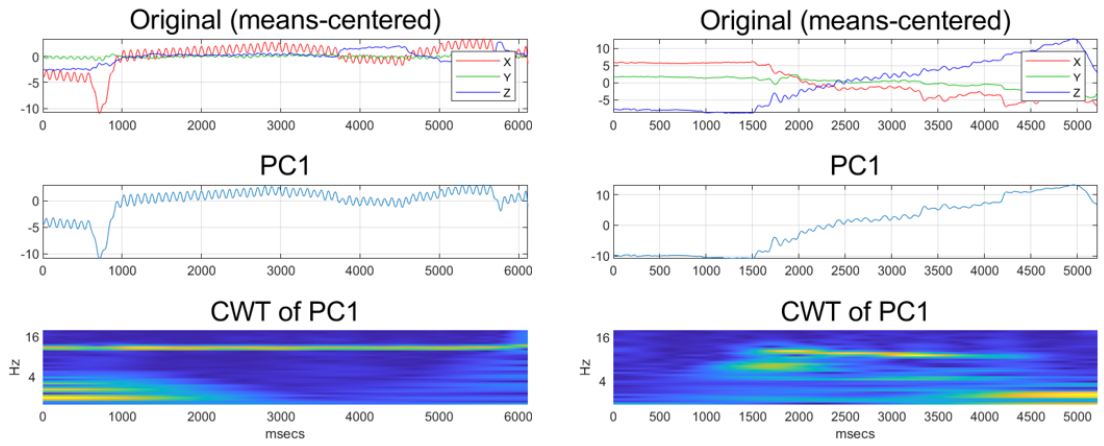


Figure 1: *Graphs associated with one example trial of Participant A (left) and Participant B (right). Top: original (means-centered) trajectory of the tongue tip during a single trial. Middle: the first principal component of the original trajectory. Bottom: a scalogram showing absolute CWT values by time and frequency. Brighter areas represent significant movement.*

Example 2. Participant B was a 65-year old male, likewise diagnosed with idiopathic PD five years prior to being included in the study (MDS-UPDRSIII: 21 points; TD phenotype). In the real-time visual feedback task, he showed action (kinetic or postural) tremor of the tongue tip at 8.2 Hz (SD = 0.8) with an RMS amplitude of 0.3 mm (SD = 0.1; see Figure 1, right). He showed no tremor during the speech production or sustained phonation tasks.

Discussion. Tremor assessment of the full dataset, including control speakers, is ongoing. Current examples suggest that EMA sensors could potentially constitute a reliable method to detect not only resting tongue tremor but also action tremor in different tasks. This adds to our prior study (Jacobi et al. 2020) which demonstrated only rest tremor in frequencies of 3-4 Hz. Detecting different types of lingual tremor may have important clinical implications, as these may relate to swallowing and speech motor control difficulties in IwPD (Robbins, Logemann, and Kirshner 1986).

References.

Bhatia, K. P., P. Bain, N. Bajaj, R. J. Elble, M. Hallett, E. D. Louis, J. Raethjen, M. Stamelou, C. M. Testa, G. Deuschl, and the Tremor Task Force of the International Parkinson and Movement Disorder Society (2018). "Consensus Statement on the classification of tremors. from the task force on tremor of the International Parkinson and Movement Disorder Society". In: *Movement Disorders*.

Chan, P.Y., Z.M. Ripin, S.A. Halim, W. N. Arifin, A. S. Yahya, G. B. Eow, K. Tan, J. Y. Hor, and C. K. Wong (2022). "Motion characteristics of subclinical tremors in Parkinson's disease and normal subjects." In: *Scientific Reports volume 12*.

Goetz, C. G., B. C. Tilley, S. R. Shaftman, G. T Stebins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, et al. (2008). "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results". In: *Movement Disorders*.

Hunkers, C. J. and J. H. Abbs (1990). "Uniform frequency of parkinsonian resting tremor in the lips, jaw, tongue, and index finger". In: *Movement Disorders*.

Jacobi, J., T. Rebernik, R. Jonkers, B. Maassen, M. Proctor, and M. Wieling (2020). "Characterizing tongue tremor in Parkinson's disease using EMA". In: *Proceedings of the 12th International Seminar on Speech Production*.

Lilly, J. M. (2017). "Element analysis: a wavelet-based method for analysing time-localized events in noisy time series". In: *Proceedings of the Royal Society A*.

Rebernik, T., J. Jacobi, R. Jonkers, A. Noiray, and M. Wieling (2021). "A review of data collection practices using electromagnetic articulography". In: *Laboratory Phonology*.

Rebernik, T., J. Jacobi, M. Tiede, and M. Wieling (2021). "Accuracy assessment of two electromagnetic articulographs: Northern digital inc. wave and northern digital inc. vox". In: *Journal of Speech, Language, and Hearing Research*.

Robbins, J. A., J. A. Logemann, and J. S. Kirshner (1986). "Swallowing and Speech Production in Parkinson's Disease". In: *Annals of Neurology*.

Stebbins, G. T., C. G. Goetz, D. J. Burn, J. Jankovic, T. K. Khoo, and B. C. Tilley (2013). "How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: Comparison with the unified Parkinson's disease rating scale". In: *Movement Disorders*.

Speech Rhythm as a Coordinative System Stabilizing Speech Production in Auditory Feedback Perturbations

Jinyu LI[1], Leonardo Lancia[2]

[1]Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle)
[2]Laboratoire Parole et Langage (CNRS & Aix-Marseille Université)
jinyu.li@sorbonne-nouvelle.fr, leonardo.lancia.1@univ-amu.fr

Introduction. In speech production, acoustic energy exhibits variations at both the syllable rate and the rate of prominent syllables (promSyllable). That is, specific syllables within an utterance may possess acoustic prominence, often associated with factors such as lexical stress. According to [1], grouping of elements (e.g., syllables) occurs in all languages, but its strength may vary across languages depending on their phonology. Research has demonstrated that acoustic physical behaviors at distinct time scales (e.g., syllable rate rhythm and promSyllable rate rhythm) are interconnected in a stable state specific to the spoken language and the conditions of speech production [2]. For instance, the temporal relationship between these two rhythmic components exhibits greater stability in languages characterized as so-called "stress-timed" compared to those termed "syllable-timed" [3]. Furthermore, a speech with highly perceptual rhythmicity has been revealed to be correlated with heightened temporal coordination between these two components [4]. Our hypothesis posits that, under the condition of delayed auditory feedback (DAF), speakers employ a mechanism to stabilize their speech production by reinforcing the temporal coordination between the rhythm at syllable rate and the rhythm at promSyllable rate. Moreover, this reinforcement could be attributed to a higher cohesion between grouped syllables.

Experiment. To investigate our hypothesis, we conducted a speech production experiment involving DAF. 21 native French female speakers participated in the experiment, with each individual tasked with repeating three five-syllable sentences 96 times. Each trial consisted of one repetition of the three sentences in a random order without interruptions. The initial six trials are considered baseline trials as no DAF was applied. In each subsequent trial, the DAF was randomly set at 0, 60, or 120 ms, with each DAF level applied an equal number of trials throughout the entire experiment.

Methodology. The method used to evaluate the temporal coordination between syllable rhythm and promSyllable rhythm is based on the analysis of amplitude modulations (AMs) in the acoustic signal of speech. The signal is band-pass filtered to separate changes due to syllable production from the slower changes resulting from the production of the most promSyllables. It is worth noting that the recorded speech was generated in the context of DAF, which means that the speech rate may present a great variability. Therefore, we adapted the cutoff frequencies of the filters to each recording, i.e., to each experimental trial. This process generates two oscillatory signals: syllAM (syllabic amplitude modulations) and accAM (prominence amplitude modulations). The first signal should contain peaks mainly at the level of the syllabic nuclei, while the peaks of the second signal should be related to the presence of the more promSyllables (see an example in Figure 1). Due to the oscillatory nature of the signals obtained, their instantaneous phase was calculated, indicating at each instant the position of each signal in its cycle with angular values ranging from 0 to 2 π. The instantaneous phase values varying over time made it possible to measure the degree of coordination (inversely related to the variability of their temporal relationship) between syllAM and accAM in different temporal windows using the calculation of the Phase Locking Value (PLV). This measure quantifies the variability of the lag between the two signals (relative to the duration of their cycles) in each temporal window considered.

Results. Our AM analysis effectively distinguished between two rhythmic components (i.e., extracted syllAM and accAM with different numbers of peaks) across most conditions. Furthermore, we identified five syllable peaks in most sentence repetitions, correlating with the phonological syllable numbers. In these sentences with five syllable peaks identified, various numbers of prominent syllable (promSyllable) peaks were identified (see the four panels above in Figure 2). Most sentences contained either three or four promSyllable peaks. The effect of DAF on the PLV varied according to the identified promSyllable peak numbers (see the four panels below in Figure 2). On one hand, for sentences containing three promSyllable peaks, DAF tended to increase the PLV between syllAM and accAM, especially when the DAF was at 120 ms. On the other hand, more sentences containing four or five promSyllable peaks were observed in the context of DAF. This implies that the presence of DAF promoted speakers to produce a staccato-style speech. That is, speakers tended to assign the same prominence to each syllable under the influence of DAF. The staccato-style sentences had initially high PLV due to the similarity between the syllAM and accAM of these sentences. Thus, DAF has a smaller effect on the PLV of these sentences.

Discussion. Establishing stable temporal coordination between syllable rhythm and promSyllable rhythm is imperative for maintaining production stability under DAF conditions. The heightened coordination observed between syllable and promSyllable rhythms under DAF conditions may stem from increased cohesion among grouped syllables. That is, when subjected to temporal perturbations caused by DAF, a higher organizational level (e.g., accAM) might be reinforced to simplify the control complexity at the lower levels (e.g., syllAM). This study implies that speech rhythm patterns emerge from the interactions of diverse processes, encompassing sensorimotor control and accentuation.

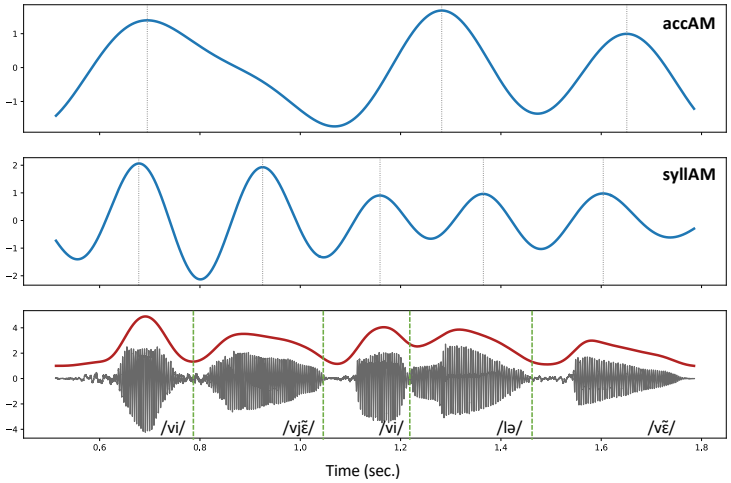


Figure 1: *Example of the extracted syllAM and accAM for the sentence /vivjɛ̃ vi lə vɛ̃/. The green vertical lines indicate phonological syllable boundaries. The grey vertical lines indicate the time points of peaks identified by our algorithm in each AM.*

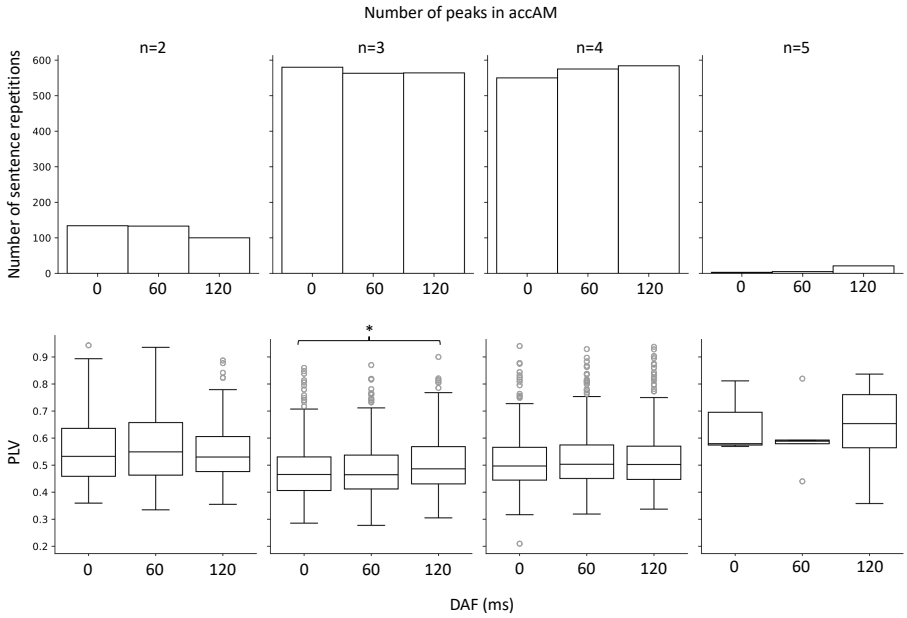


Figure 2 *Distribution of the number of repetitions containing various numbers of prominent syllable peaks in accAM for the sentences with* ***five syllable peaks*** *in syllAM (four panels above) and the PLV at each level of DAF in the corresponding sentences (four panels below).*

References

[1] Auer, P. (1993). Is a rhythm-based typology possible. *A study of the role of prosody in phonological typology*.

[2] Cummins, F., & Port R. (1998). Rhythmic Constraints on Stress Timing in English. *Journal of Phonetics* 26(2):145–71.

[3] Lancia, L., Krasovitsky G., & Stuntebeck F. (2019). Coordinative Patterns Underlying Cross-Linguistic Rhythmic Differences. *Journal of Phonetics* 72:66–80.

[4] Leong, V., & Goswami U. (2014). Assessment of Rhythmic Entrainment at Multiple Timescales in Dyslexia: Evidence for Disruption to Syllable Timing. *Hearing Research* 308:141–61.

Articulatory timing in Hindi CV sequences

Shihao Du[1], Indranil Dutta[2], Adamantios I. Gafos[1]

[1]Univeristät Potsdam, Potsdam, Germany, [2]Jadavpur University, Kolkata, India

shihao.du@uni-potsdam.de, indranildutta.lnl@jadavpuruniversity.in, gafos@uni-potsdam.de

Introduction. In stop-vowel sequences of Hindi, we asked how phonation and place of articulation of a consonant as well as vowel quality influence the timing of the consonantal and vocalic oral gestures. Intervals delineated by landmarks on CV sequences have been examined in works that assess the extent to which inter-segmental coordination can be expressed in terms of synchronicity relations among landmarks (Gafos 2002). For instance, Kramer et al. (2023) report the mean and standard deviation of four intervals (C-onset to V-onset, V-onset to C-target, C-target to V-target, V-target to C-offset) on the basis of eight word-initial CV sequences in American English and Mandarin, where the initial consonant is either /b/ or /m/ and the vowel is either low back /ɑ/ or high front /i/. Out of the four intervals examined in Kramer et al. (2023), V-target to C-offset was the one with a mean closest to zero (implying near synchronicity of the two landmarks); see also Shaw & Chen (2019) and Durvasula & Wang (2023). In the current work on Hindi, we adopt the V-target to C-offset interval to quantify CV timing and examine how consonant phonation, place of articulation, and vowel quality modulate this interval.

Methods. Electromagnetic articulography data were collected from 2 native male speakers of Hindi aged at 22 and 23. The speakers produced 63 target words beginning with CV sequences where the consonant was either /b/, /p/, /bʱ/, /d/, /t/, /dʱ/, or /tʰ/ (/pʰ/ is not included because in Hindi it underwent fricativization) and the vowel was one of / iː, ɪ, uː, ʊ, eː, e, oː, o, aː/. Each target word was repeated 10 times by each speaker. The Carstens AG501 device was used to record movements at a sampling rate of 1250 Hz. A linear-mixed effects model was fitted to the Hindi data with the synchrony measure as the dependent variable and consonant voicing (voiced vs. voiceless), aspiration (aspirated vs. un-aspirated), place (dentals vs. labials), vowel height (high vs. low vs. mid), vowel frontness / roundness (back / rounded vs. non-back / unrounded), and vowel length (long vs. short) as fixed effects (all sum-coded). Random intercepts for speakers and items were also included.

Results. Figure 1 presents density plots of the V-target to C-offset interval as a function of the six fixed effects (consonant voicing, aspiration, place, vowel height, frontness / backness, and length). The model had an intercept of 3.14 ms, indicating that the vowel target occurs on average approximately 3 ms before the consonant offset. An ANOVA test applied to the linear-mixed effects model revealed that consonant place, vowel height and frontness / backness had significant effects on the synchrony measure (p-value < 0.0001 for all three; F-value = 24.50, 24.83, and 17.01 respectively), whereas the effects of consonant voicing, aspiration, and vowel length did not reach significance (p-value = 0.17, 0.56, and 0.20 respectively; F-value = 1.86, 0.33, and 1.65 respectively). For the significant effects post-hoc pairwise comparisons were implemented using the R package EMMEANS (Lenth et al. 2023). For consonant place, the comparisons indicate that the two landmarks are 12.9 ms farther apart when C place is dental versus labial. In terms of vowel height, the synchrony measure was 14.8 ms shorter in high compared to mid vowels and 26.2 ms shorter in low compared to mid vowels; with regard to frontness/backness, back rounded vowels had 10.9 ms longer lag than non-back unrounded vowels.

These results raise the question why CV timing, as quantified by the interval from V-target to C-offset, is more sensitive to vowel quality (vowel height and frontness) than to consonant phonation (voicing and aspiration). A possible explanation is that CV timing is determined by the kinematics (displacement, peak velocity, stiffness, etc.) of the relevant movements in the CV transition which have been shown to be more susceptible to the influences of vowel-related than consonant-related properties. For instance, early studies on English CV sequences (Ostry et al. 1983, Löfqvist and Gracco 1997) reported robust vocalic effects on the consonant's kinematics, while the effects of consonant voicing on these kinematics appear to be place-specific or not consistent across subjects. Thus, Löfqvist and Gracco (1997) reported no consistent voicing effect in labial consonant-initial CVs, whereas Ostry et al. (1983) reported such an effect on C displacement and peak velocity in the opening and closing movements for velar consonant-initial CVs. To assess if and how these results on differential effects of consonant and vowel properties on the consonant's kinematics also extend to Hindi's more elaborate system of contrasts, we fitted the model described in the Methods section to our data with six kinematic measures from the consonantal gesture as the dependent variable: displacement, peak velocity, and stiffness of the closing and opening movements. In **Table 1** below, we summarize the significant effects for each kinematic measure grouped by whether they are related to the consonant or the vowel. It can be seen that while the kinematics of the consonantal closing movement are modulated by both consonant and vowel-related factors, those of the opening movement are almost exclusively vowel-sensitive and immune to consonant phonation. Therefore, effects related to

consonant phonation (i.e., voicing and aspiration) on gestural kinematics are not only dwarfed by vocalic effects in terms of the number (3 significant aspiration / voicing effects vs. 8 significant height / frontness effects), but are also highly localized on the consonantal closing movement as opposed to the opening movement. Since CV timing mainly concerns the transition between C and V, which mostly encompasses the C opening and V movement, the lack of consonantal effects on the kinematics of the consonantal opening movement may be the reason why CV timing is insensitive to consonant phonation as revealed by our results on CV landmark synchronicity shown above.

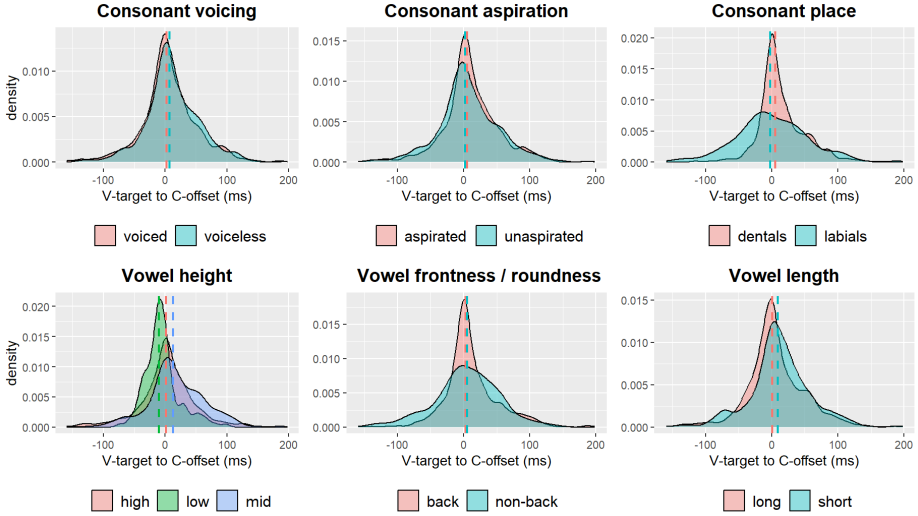


Figure 1: *Distribution of the V-target to C-offset interval across subjects as a function of consonant voicing, aspiration, place, vowel height, frontness / roundness, and length. Vertical lines are the medians in each group.*

C movement	Kinematic measure	Consonant-related	Vowel-related
Closing movement	displacement	Place***, aspiration***	Height***, frontness*
	peak velocity	Place***, aspiration***	Height***
	stiffness	Place***, voicing**	Frontness***
Opening movement	displacement	/	Height***, frontness**
	peak velocity	/	Height ***
	stiffness	Place**	Frontness***

Table 1: *Significant effects of consonant and vowel-related factors on gestural kinematics of the consonantal closing and opening movements. Forward slashes denote the absence of significant effects. Asterisks denote the level of statistical significance for each effect in terms of p-value.*

Conclusion. Vowel height/frontness and C place of articulation exert significant effects on landmark synchrony as measured by the interval from V-target to C-offset, whereas C phonation of the initial stop has no significant effect. We sought to explain this finding by demonstrating, in an extension of earlier work on English, that while vowel quality significantly affects movements towards and away from the C constriction, effects of C phonation are confined to the kinematics of the closing movement alone. That is, such effects are absent in the opening movement, which is the one directly involved in CV gestural transition. This may then explain the presence of vowel quality effects and the absence of consonant phonation effects in CV timing.

References

Durvasula, K., & Wang, Y. (2023). Revisiting CV timing with a new technique to identify inter-gestural proportional timing. *Conference Proceedings of the 20th International Congress of Phonetic Sciences*, 2284–2288. https://drive.google.com/file/d/15U2l2y4_-9lyZAgmiccQYXYj9zBi_CAu/view

Gafos, A. I. (2002). A grammar of gestural coordination. *Natural Language & Linguistic Theory*, 20, 269–337.

Kramer, B. M., Stern, M. C., Wang, Y., Liu, Y., & Shaw, J. A. (2023). Synchrony and stability of articulatory landmarks in English and Mandarin CV sequences. *Conference Proceedings of the 20th International Congress of Phonetic Sciences*, 1022–1026. https://drive.google.com/file/d/15U2l2y4_-9lyZAgmiccQYXYj9zBi_CAu/view

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2023). *emmeans: Estimated marginal means, aka least-squares means (Version 1.8.9)* [Computer software]. Retrieved from: https://cran.r-project.org/web/packages/emmeans/index.html

Löfqvist, A., & Gracco, V. L. (1997). Lip and jaw kinematics in bilabial stop consonant production. *Journal of Speech, Language, and Hearing Research*, *40*(4), 877–893. https://doi.org/10.1044/jslhr.4004.877

Ostry, D. J., Keller, E., & Parush, A. (1983). Similarities in the control of the speech articulators and the limbs: kinematics of tongue dorsum movement in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(4), 622–636. https://doi.org/10.1037/0096-1523.9.4.622

Shaw, J. A., & Chen, W. (2019). Spatially conditioned speech timing: evidence and implications. *Frontiers in Psychology*, 10, 2726. https://doi.org/10.3389/fpsyg.2019.02726

What's in a name? Production (and Perception) of Difficult to Pronounce Names in Academic Settings

Michelle Middaugh Cifuentes[1], Daniel Pape[2]

[1] University of Alberta
[2]McMaster University
mmiddaug@ualberta.ca, paped@mcmaster.ca

Introduction. Proper names encapsulate one's identity and are the primary identifiers of the self. Previous studies found that the correct pronunciation of proper names, especially the given name, is important to many individuals (see e.g. Zhang & Noels, 2021). However, for individuals who possess non-Western names, incorrect pronunciation of their given name is something they face daily in their interactions with others. On the other side, native English speakers unfamiliar with different languages, especially non-Western languages, may of course struggle to pronounce foreign names on different phonetic levels, e.g. of course with respect to non-native phonemes, but also with stress placement or phonotactics. The impact this has on different individuals varies, with some who feel that the mispronunciation of their name is disrespectful to their culture and identity (Payne et al. 2016), and some who do not mind (Zhang & Noels, 2021). In University settings, it is frequent that students with difficult to pronounce[1] names have their names mispronounced, and many choose to adopt a Westernized name to avoid that mispronunciation. It is therefore important to determine the impacts that proper name mispronunciation on different levels has on students with foreign names, and to determine how this impact extends to and shapes their academic pursuits, and thus furthermore their future career and life (see e.g. Laham et al. 2012 for the effects of name pronunciation on career hierarchy).

Aims of the Study: This study aims to examine (1) how English native speakers adapt to the phonetic complexity of difficult to pronounce (non-Western) names when attempting to faithfully reproduce these names, (2) what the perceptions are that these native English speakers have of names from different language background with respect to likeability and ease of pronunciation, and (3) the impact that frequent mispronunciation of names actually has on the affected individuals, i.e. students with difficult to pronounce (non-Western) names. Here we present preliminary results of this study.

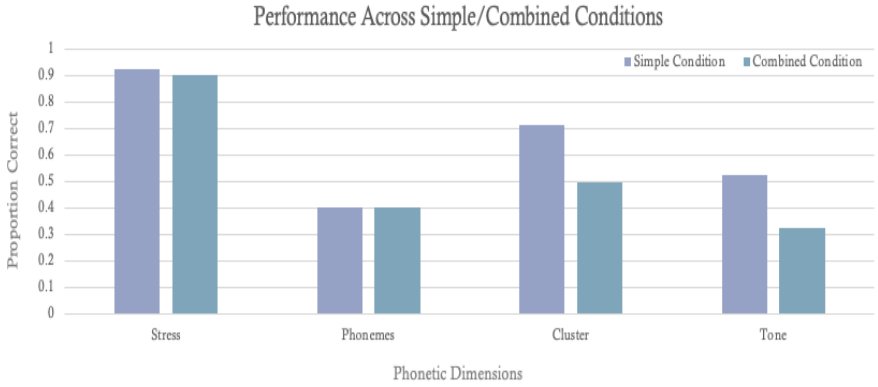
Methods.
Experiment 1 encompasses a combined perception/production study that tested 22 native English participants for their ability to perceive and then re-produce (i.e. imitate) novel name stimuli from 5 different language backgrounds (German, Korean, Mandarin, Spanish, Gujarati). These 50 stimuli varied on 5 different phonetic levels: number of syllables, stress, foreign phonemes, complex consonant clusters and lexical tone. Names were recorded by native speakers of these languages (recorded in native language context) and cross-spliced into an English sentence frame (e.g. "Hi. My name is __." Each condition was presented separately (i.e. one phonetic parameter varied: *simple condition*) or all phonetic parameters combined within one name stimulus (*complex condition*). The aim of including the complex condition was to determine the hierarchy of phonetic parameters used by listeners with respect to foreign name reproduction, or in other words which dimensions would be truthfully reproduced while other dimensions would possibly be ignored by listeners/speakers. Thus, we aimed to test the cue trading of different phonetic dimensions when English participants attempted to perform the name imitation/reproduction task. Following that imitation task, the English participants also quantitatively rated all presented names on two Likert scales, one for name likability and one for ease of pronunciation.
Experiment 2 used qualitative interviews with 10 scripted questions. Participants were students who possess difficult to pronounce names (n=7) with the aim to determine their experiences with name mispronunciation across their lifespan and within academic settings.

Results.
Experiment 1: A reproduction of one of the phonetic dimensions was considered successful when two raters (first author and native speaker of the language in question) considered the participant's effort to faithfully reproduce as meaningful and substantial. For the preliminary results, as can be seen in Figure 1, we found that the areas of pronunciation difficulty for native English speakers were mostly in the areas of foreign phonemes and lexical tone, but interestingly the number of syllables and thus the length of the name (within the limits tested here, i.e. three versus four syllables) does not seem to play a role for accuracy here (see Figure 2 for a comparison). Furthermore, participants' ability to accurately replicate complex consonant clusters and attempts to lexical tone was impacted the most in the *complex condition*, indicating that participants will favor correct stress allocation and reproduction of foreign phonemes here, or rather their attempts at these categories. This is especially true for the foreign phoneme dimension, which saw the poorest performance from

[1] in the sense of: difficult to pronounce for English speakers in an English speaking University setting

participants overall and suggests that participants are using most of their cognitive resources to work around the pronunciation of these foreign phonemes when they are presented simultaneously with other phonetic dimensions.

Figure 1: *Accuracy reproduction ratings for the four phonetic dimensions stress, foreign phonemes, consonant clusters and lexical tone, split by simple conditions (each phonetic parameter varied individually) versus combined (all phonetic parameters present within one name stimulus).*



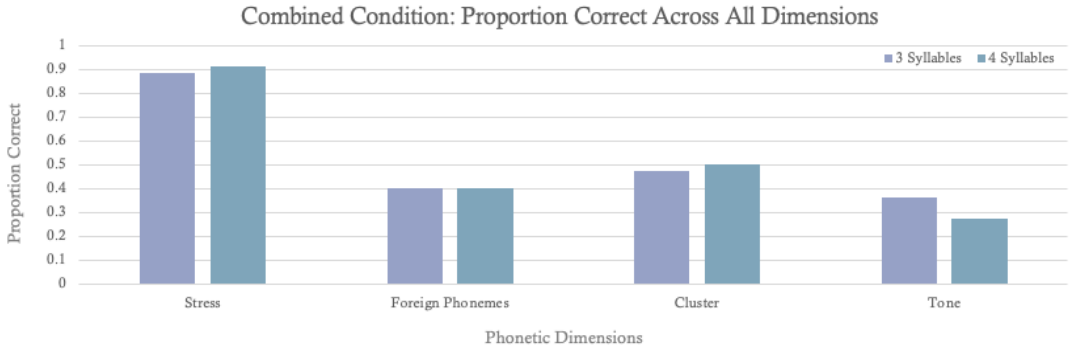


Figure 2: *Accuracy reproduction rating differences between three syllable and four syllable stimuli for the four phonetic dimensions stress, foreign phonemes, consonant clusters and lexical tones for the combined condition (all phonetic parameters present within one name stimulus).*

The results of the likeability and ease of pronunciation scales that the native English participants had to rate showed that participants considered easy to pronounce names as more likeable (based on their scale ratings), although the opposite trend is less evident for difficult to pronounce names.

Finally, we determined the participants' attitudes towards these foreign names with a name selection task. This task followed the reproduction and rating part and tasked participants to select a small subset of all novel name stimuli based on a hypothetical academic cooperation-based scenario. Participants were told to pick 5 names out of the full 50 name set (10 stimuli per language x 5 languages) for collaboration in a group assignment. The results of this name selection task showed that participants overwhelmingly chose names from the Spanish language background, but avoided names from other languages[1].

Experiment 2: While several answers and themes strongly varied (among these: positive versus negative experiences with the difficult to pronounce names, experiences in childhood versus adolescence, problems during application processes, closer versus more distant connection to their heritage culture) the main emerging theme from this part of the experiment was that students did not feel that their academic pursuit was strongly impeded by their difficult to pronounce names. However, they highly appreciated if their names *are* pronounced correctly by professors and their peers. Many participants also stated that even attempts to make an effort (to the best of the English speakers' abilities) in order to try to pronounce their names correctly made significant positive impacts and were highly appreciated.

References

Laham, S. M., Koval, P., & Alter, A. L. (2012). The name-pronunciation effect: Why people like Mr. Smith more than Mr. Colquhoun. Journal of Experimental Social Psychology, 48(3), 752–756. https://doi.org/10.1016/j.jesp.2011.12.002

Zhang, Y. S., & Noels, K. (2021). The frequency and importance of accurate heritage name pronunciation for post-secondary international students in Canada. Journal of International Students, 11(3), 608–62.

[1] It is acknowledged here that it is possible that unconscious (or conscious) bias towards particular national or ethnic groups (mirrored in the names) is influencing results, rather than the name per se. This study could not test for this effect.

Association between Speech Motor Learning and Model-based Estimates of Memory Retrieval

Thomas Wilschut[1], Katharina M. Polsterer[2], Thomas Tienkamp[2,3], Hedderik van Rijn[1], Catherine Sibert[4], Defne Abur[2]

[1]Department of Experimental Psychology, University of Groningen, [2]Center for Language and Cognition Groningen, University of Groningen, [3]Department of Oral and Maxillofacial Surgery, University Medical Centre Groningen, [4]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen

d.abur@rug.nl

Introduction. The degree of speech motor learning (the amount that motor speech commands are adjusted based on changes to sensory input) has been shown to be influenced by various cognitive factors, such as attention, declarative memory and cognitive load (e.g., Liu et al., 2018). In line with this, individuals with changes in the domains of memory function (e.g., people with Alzheimer's disease) demonstrate disruptions to speech motor control (e.g., Liu et al., 2012). Even in typical speakers, there is a wide range in memory capacities and the degrees of speech motor learning (Lametti et al., 2012, an Rijn, van Maanen & van Woudenberg, 2009). Since speech motor learning requires accessing stored motor programs, and interacts with more global mechanisms underlying memory retrieval, here we explored whether the degree of speech motor learning from an auditory perturbation task was associated with model-based estimates of memory in typical speakers.

Methods. A total of 22 typical speakers (females = 16, males = 6; aged 18 - 25), with no speech, language, hearing, or cognitive impairment, participated in the experiment. All participants passed a standard hearing screening. Participants completed two tasks: one for declarative memory retrieval (via MemoryLab; see www.memorylab.nl/en/) and one for speech motor learning.

The MemoryLab learning task consisted of a 12-minute adaptive retrieval practice session, where the learners were asked to learn specific new words. The task was completed in a quiet room. During retrieval practice, the system calculated a rate of forgetting (i.e., the speed at which the learner was forgetting the items) based on the response times and accuracy scores for each individual retrieval attempt (e.g., see van Rijn, van Maanen & van Woudenberg, 2009). After the experiment, an average rate of forgetting over all items was computed, and used as a measure of memory capacity (e.g., see Zhou et al., 2021).

The speech motor learning task was completed in a sound-attenuated booth and consisted of a gradual perturbation paradigm using prolonged words. Participants were asked to produce ‘bid’, ‘bed’ and ‘bad’ (all real words in Dutch) in a random order. They were recorded with an over-the-ear microphone (Shure MX153), and received real-time auditory feedback (amplified by 5 dB) via headphones while speaking. The experiment comprised a total of 108 trials split over four phases: baseline, ramp, hold, and after-effect. During the 24 baseline trials, the participants received unperturbed auditory feedback. Over the course of the 30 subsequent ramp trials, the first formant (F_1) in the auditory feedback was gradually increased by 1.7% per trial. For the 30 trials of the hold phase, the % F_1 increase was held constant at 50% relative to the mean F_1 of the baseline phase. During the last 24 trials constituting the after-effect phase, there was no perturbation. For each trial, the mean F_1 in Hz was measured in a window of 40 – 120 ms. The degree of speech motor learning was quantified as the mean % change in F_1 across the hold phase (i.e., the change in production during the maximal sensory difference).

To further explore the relationship between memory retrieval and speech motor learning, a model of the speech motor learning task was built using the ACT-R cognitive architecture (Anderson et al, 1997). ACT-R is a high level abstraction of whole brain cognition, and includes multiple basic modules representative of cognitive functions as well as mechanisms that govern the interaction and transfer of information between them. Hence, the ACT-R architecture was used to model both a global rate of forgetting and speech-specific learning in reference to the experimental tasks.

Results. Analyses are ongoing, but results are shown for the speakers analyzed thus far (N = 8/22). The preliminary results suggest a trend for a positive association between the rate of forgetting and the amount of speech motor learning during the speech task (R = 0.454, Figure 1). The full dataset will be integrated into the built ACT-R model framework to model the relation between these two measures as well.

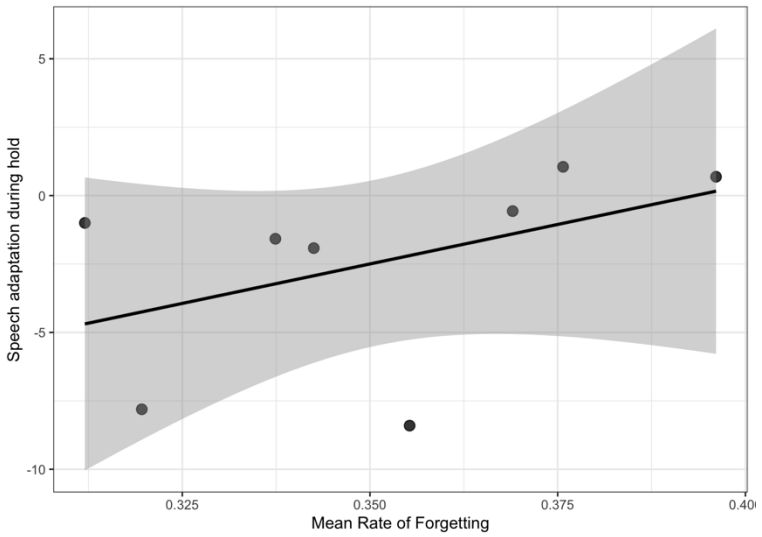


Figure 1: *The mean rate of forgetting (unitless number output from the MemoryLab task) is plotted against the speech motor learning measure (% change F_1 relative to the baseline).*

Discussion The current work assessed the relation between memory capacities (i.e., rate of forgetting during a learning task) and the amount of speech motor learning (i.e., the degree of F_1 changes in a speech perturbation task) in typical speakers. Analyses are ongoing for the data collected, but preliminary results suggest a trend for a positive relationship between the two measures. The results of this study will aid our understanding of the relationship between the degree of speech-specific motor learning and memory capabilities as measured by a word learning task. Comparing the built ACT-R model predictions to the behavior patterns of the human participants will also provide additional insight into how memory function interacts with speech.

References

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human–Computer Interaction*, *12*(4), 439-462.

Lametti, D. R., Nasir, S. M., & Ostry, D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. Journal of Neuroscience, 32(27), 9351-9358.

Liu, Y., Fan, H., Li, J., Jones, J. A., Liu, P., Zhang, B., & Liu, H. (2018). Auditory-motor control of vocal production during divided attention: behavioral and ERP correlates. Frontiers in Neuroscience, 12, 113.

Liu, H., Wang, E. Q., Metman, L. V., & Larson, C. R. (2012). Vocal responses to perturbations in voice auditory feedback in individuals with Parkinson's disease. PloS one, 7(3), e33629.

McDougle, S. D., Ivry, R. B., & Taylor, J. A. (2016). Taking aim at the cognitive side of learning in sensorimotor adaptation tasks. Trends in cognitive sciences, 20(7), 535-544.

Van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009, July). Passing the test: Improving learning gains by balancing spacing and testing effects. In Proceedings of the 9th International Conference of Cognitive Modeling (Vol. 2, No. 1, pp. 7-6).

Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of sensorimotor learning. Nature reviews neuroscience, 12(12), 739-751.

Zhou, P., Sense, F., van Rijn, H., & Stocco, A. (2021). Reflections of idiographic long-term memory characteristics in resting-state neuroimaging data. Cognition, 212, 104660.

Articulatory and vocal speaker variability in connected speech

Maria Mendes Cantoni[1], Adelino Pinheiro Silva[2]

[1]*Federal University of Minas Gerais*
[2]*Police Academy of Minas Gerais*
mmcantoni@gmail.com, adelinocpp@gmail.com

Introduction.
Speech is subject to multiple sources of variation and each occurrence of a sound in a language is a particular and unique instance. Speaker variability is considered an obstacle to the description of the sound system of a language and the study of other types of sociophonetic variation, to the extent that it affects the realization of speech sounds. Modelling speaker variability is important not only to understand the speech process in a linguistic perspective, but also for psychological, clinical and biomechanical studies and to advance applications that require speaker identification, such as forensic phonetics and authentication systems (Brunner 2009). Two types of speaker variability are classically devised (Kilbourn-Ceron and Goldrick 2021). Between-speaker variability can be defined as variation in speech due to different vocal tracts or due to different motor routines displayed by individuals. Within-speaker variability refers to variation due to slight differences on how speech movements are actually implemented by the same individual. An individual can be recognized by his or her speech and voice (Kreiman and Sidtis 2011), yet the role of different components of the vocal tract in speaker identification is not clear (Lee, Keating, and Kreiman 2019). Furthermore, aiming at a finer control of phonological and prosodic conditions, most studies used read speech to evaluate speaker variability and only a few used connected speech (Lee and Kreiman 2022). In this study we address the two types of speaker variability, with the aim to untangle the role of articulatory structures and voice in speaker identification in connected speech. We intend to answer the following questions: how much speaker variation is due to articulation differences and how much is due to voice differences? Which acoustic measures are more robust for speaker identification in connected speech?

Methods. The database of spoken Brazilian Portuguese CEFALA-1 (Neto, Silva, and Yehia 2019) was used. A sample of 18 speakers (10 female and 8 male) from the same dialect was randomly selected and the portion containing connected speech (average length of 2 minutes) was evaluated. In each audio, all vowels were manually segmented and labeled by trained researchers. The tokens were coded for the phonological factors: vowel quality, preceding and following sound context, stress degree, number of syllables of the word, and syllable structure. Only the vowels were used to evaluate variability, since they carry both phonatory and articulatory components and the open vocal tract presents a less complex mapping between articulation and acoustics. A total of 5614 vowels and diphthongs were segmented in the recordings and 47 acoustic measures classically used in speech and voice studies (Garellek 2022; Kreiman and Sidtis 2011) were performed, belonging to two sets: articulatory (related to vocal tract resonances: duration, COG, the mean, slope and concavity of the 8 first formants) or phonatory (related to harmonicity: HNR, intensity, mean, slope and concavity of F0, H1*-H2*, H2*-H4*, LTF). In order to evaluate the research questions, a statistical modelling procedure was used that could extract away phonologically predictable vowel characteristics. First, each acoustic variable was fit to a generalized linear model with phonological factors as predictors. The residuals of the models were used as data tokens, since they would carry more information related to speakers than the original measures. Then, the data was z-score normalized and the Euclidean distances between the tokens were calculated. The data was divided into two sets: training (n = 3921, with a mean of 218 vowels per speaker) and test (n = 1693, with a mean of 94 vowels per speaker). A logistic regression classifier was used to assign each token to the same speaker or different speaker class.

Results. The classifier reached an accuracy of 100% with equal error rate of 0% and Log Likelihood Ratio Cost (C_{LLR}) of 0.006 (see Brümmer and Du Preez 2006) in a threshold of 0.85 log likelihood ratio. Figure 1 shows the separation of the data tokens in residual space, after the classification procedure. The acoustic variables that were more relevant to distinguish between speakers were: time duration; bias of the third, fourth, sixth and seventh formants; bias of COG; bias and concavity of H2*-H4*. Those acoustic variables consist of 5 articulatory and 2 vocal variables. Together, they explained 50% of the variance in the data set.

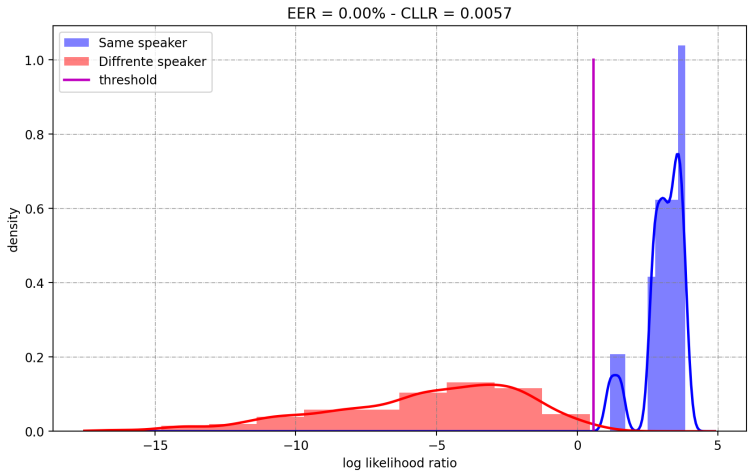


Figure 1: *Separation of data tokens in residual space between the classes "same speaker" and "different speaker", after a logistic regression classification procedure based on the Euclidean distance*

Discussion. Even in a small scale data set, we were able to evaluate how different vocal tract structures affect speech individual variation. As expected, between-speaker variation was greater than within-speaker variation. Both articulation and voice contribute to the variability found in between speakers' speech production, with a numeric prevalence of articulatory over vocal characteristics. Thus vocal tract resonances, despite conveying information such as vowel quality, are relevant to explain speaker variability, and it can be expected that combining parameters for articulation and voice would allow for better performance in tasks involving speaker identification. Our results also show that the same variables relevant to between-speaker classification play a role on the within-speaker classification. It means that the patterns of variation of a single speaker may not differ in nature from the patterns found to differentiate speakers, and that variation may arise not only from anatomical differences, but from the means speakers use their vocal tract. The present results refer to a group of speakers from the same dialect. In a future work, we intend to use the same method to further understand whether the acoustic variation that explains speaker variability contributes to the separation of speakers from different dialects.

References.

Brümmer, Niko and Johan Du Preez (2006). "Application-independent evaluation of speaker detection". In: *Computer Speech & Language* 20.2-3, pp. 230–275.

Brunner, Elizabeth Gentry (May 2009). "The Study of Variation from Two Perspectives". en. In: *Language and Linguistics Compass* 3.3, pp. 734–750. DOI: 10.1111/j.1749-818X.2009.00137.x. URL: https://compass.onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2009.00137.x (visited on 12/20/2023).

Garellek, Marc (Sept. 2022). "Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality". en. In: *Journal of Phonetics* 94, p. 101155. DOI: 10.1016/j.wocn.2022.101155. URL: https://linkinghub.elsevier.com/retrieve/pii/S0095447022000304 (visited on 12/20/2023).

Kilbourn-Ceron, Oriana and Matthew Goldrick (Aug. 2021). "Variable pronunciations reveal dynamic intra-speaker variation in speech planning". en. In: *Psychonomic Bulletin & Review* 28.4, pp. 1365–1380. DOI: 10.3758/s13423-021-01886-0. URL: https://link.springer.com/10.3758/s13423-021-01886-0 (visited on 12/20/2023).

Kreiman, Jody and Diana Sidtis (2011). *Foundations of voice studies: an interdisciplinary approach to voice production and perception*. Malden, MA: Wiley-Blackwell.

Lee, Yoonjeong, Patricia Keating, and Jody Kreiman (Sept. 2019). "Acoustic voice variation within and between speakers". en. In: *The Journal of the Acoustical Society of America* 146.3, pp. 1568–1579. DOI: 10.1121/1.5125134. URL: https://pubs.aip.org/jasa/article/146/3/1568/993422/Acoustic-voice-variation-within-and-between (visited on 12/20/2023).

Lee, Yoonjeong and Jody Kreiman (May 2022). "Acoustic voice variation in spontaneous speech". en. In: *The Journal of the Acoustical Society of America* 151.5, pp. 3462–3472. DOI: 10.1121/10.0011471. URL: https://pubs.aip.org/jasa/article/151/5/3462/2839404/Acoustic-voice-variation-in-spontaneous-speech (visited on 12/21/2023).

Neto, Arlindo Follador, Adelino Silva, and Hani Yehia (2019). "Corpus CEFALA-1: Base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia / Corpus CEFALA-1: Audiovisual Database of Speakers for Biometric, Phonetic and Phonology Studies". In: *REVISTA DE ESTUDOS DA LINGUAGEM* 27.1, pp. 191–212. DOI: 10.17851/2237-2083.27.1.191-212. URL: http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/13378.

Beyond speech production: sensorimotor contribution to native and non-native phoneme perception

Tzuyi Tseng[1], Jennifer Krzonowski[1], Alice C. Roy[1], Claudio Brozzoli[2*], Véronique Boulenger[1*]*

[1]DDL, CNRS/Université Lyon 2
[2]ImpAct team, Centre de Recherche en Neurosciences de Lyon, INSERM, Lyon, France
*equal contributions

tzu-yi.tseng@cnrs.fr, jennifer.krzonowski@cnrs.fr, alice.roy@cnrs.fr, claudio.brozzoli@inserm.fr, veronique.boulenger@cnrs.fr

Introduction. Embodied theories of cognition consider language as grounded in the sensorimotor system. Converging evidence shows that speech perception induces activations of sensorimotor brain areas that are normally involved in speech production (see Skipper *et al*., 2017 for a review). This motor activity was found to be somatotopically organized depending on the place of articulation of phonemes: listening to bilabial and dental consonants activates the cortical motor representations of the lips and the tongue, respectively (Pulvermüller *et al.,* 2006). Other studies however failed to replicate this somatotopic mapping of motor cortex during (native) speech perception (Arsenault & Buchsbaum, 2016), or showed (pre)motor recruitment only when speech was degraded but still identifiable (D'Ausilio *et al,.*2012; Du *et al*., 2014; Osnes *et al.,* 2011). In parallel, the involvement of the (pre)motor cortices has also been reported in non-native phoneme perception (Wilson & Iacoboni, 2006; Schmitz *et al*., 2019). The current neuroimaging study aims to further investigate how sensorimotor regions are activated as a function of the phonological distance between native and non-native phonemes under degraded or optimal perceptual conditions.

Methods. Twenty-four monolingual French right-handed adults (17 females, mean age = 24.8 ± 3 years old) participated in a combined behavioral and functional magnetic resonance imaging (fMRI) study. Their brain activity was recorded while listening to triplets of identical consonant-vowel (CV) syllables, produced by a native French-Mandarin Chinese bilingual female speaker. The syllables embedded either a native French consonant (/p/, /t/, /ʃ/ or /ʁ/) or a non-native Mandarin Chinese consonant (/pʰ/, /tʰ/, /ʂ/ or /x/), always followed by the same acoustically vowel /a/. Only a single phonetic feature differentiates between French and Chinese consonants in three of the pairs (aspiration for the labial and dental plosives /p-pʰ/ and /t-tʰ/, respectively, and tongue retroflection for the fricatives /ʃ-ʂ/). By contrast, the fricatives /ʁ-x/ differ by two phonetic features (voicing and place of articulation). Each syllable triplet was presented 18 times during the experiment. All speech stimuli (80 dB SPL) were randomized and half of them were masked with pink noise (72 dB SPL). Participants performed an Alternative Forced-Choice (2AFC) identification task in which they indicated, as accurately and rapidly as possible, whether the heard consonants were French or from a foreign language by a left-hand button press. A sparse fMRI acquisition sequence (i.e. three silent TRs (repetition times) interspersed with three acquisition TRs) was designed to reduce the impact of the high-frequency scanner noise over the perception of the auditory stimuli and behavioral motor response. For the 2AFC task, participants' accuracy (% of correct consonant identification) was analyzed with a three-way repeated-measures ANOVA including language, noise and consonant as within-subject factors, and participants as a random variable. Imaging data were pre-processed with the fMRIprep pipeline (Esteban *et al*., 2019) and univariate analyses were conducted with Nilearn packages in Python (Abraham *et al.,* 2014). General linear models (GLMs) were applied for denoising procedure (Caballero-Gaudes & Reynolds, 2017).

Results. Behavioral results (**Figure 1a)** reveal main effects of noise and language ($p_s < .001$), with better performance for intact than for noisy, and for native than for non-native consonants. The main effect of consonant is also significant ($p < .001$), the non-native retroflex /ʂ/ being the most difficult consonant to identify. Interestingly, significant noise × language and noise × consonant ($p_s < .01$) interactions are found. Noise affects consonant identification more for the native than for the non-native language. This is particularly observed for the noisy native plosive /p/ and fricative /ʃ/, and to a lesser extent for noisy /t/, / which lead to lower performance than in the intact condition and than their non-native counterparts /pʰ/, /ʂ/ and / tʰ/, respectively. Regarding non-native consonants, whereas participants tend to improve for noisy relative to intact /ʂ/, their performance diminishes for /x/ in noise. Preliminary fMRI univariate analyses suggest significant activation in the bilateral (pre)motor cortex and superior temporal cortex when identifying intact and noisy native consonants compared to baseline (family-wise error FWE $p < .05$). For non-native consonants, activations in the bilateral superior temporal gyrus and right pre- and postcentral gyri are significant compared to baseline (FWE $p < .05$) in the noisy but not the intact condition. Direct contrasts between native and non-native languages (uncorrected, $p < .001$), irrespective of noise, suggest stronger left-lateralized temporal, precentral and middle frontal activations for native phonemes, whereas right middle temporal, pre/postcentral, frontal and parietal regions are more strongly activated for non-native consonants. In the intact condition, significant activations are found mostly in the left middle frontal,

supramarginal and superior temporal gyri for native *vs* non-native consonants (**Figure 1b**), while the right supramarginal gyrus and frontal cortex are activated for non-native *vs* native consonants (**Figure 1c**). Under the noisy condition, most significant activations occur in the left inferior frontal cortex and anterior superior temporal gyrus for native stimuli (**Figure 1d**), but in the right lateral occipital lobe for the non-native *vs* native contrast (**Figure 1e**).

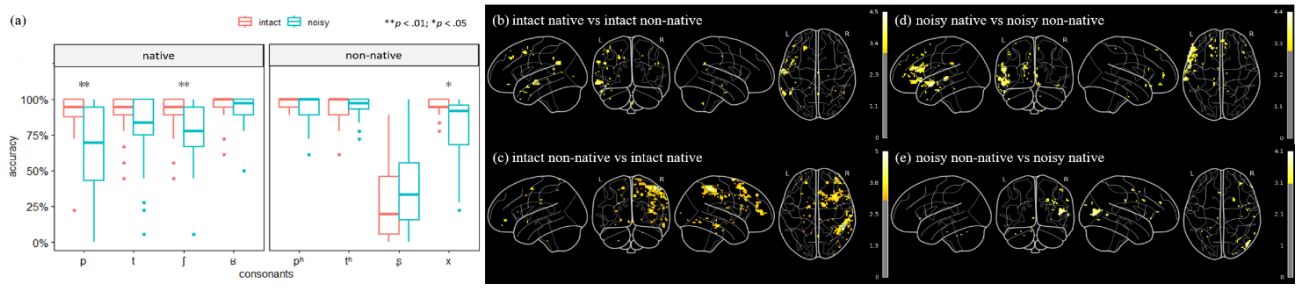


Figure 1: Mean accuracy (%) for identification of native and non-native consonants in the intact and noisy conditions (a). Whole brain activation (uncorrected p < .001) for the direct contrasts between native and non-native consonants in the intact (b and c) and noisy conditions (d and e).

Discussion. French adults exhibited good identification performance for all intact consonants from the native and non-native languages, except for the non-native retroflex fricative /ʂ/. This retroflex consonant, which differs from the native fricative by one phonetic feature that is not part of the native repertoire, was also difficult to identify in noise. Despite aspirated plosives also only differ by one acoustic feature from native unaspirated consonants, participants were able to identify them easily. Since aspiration exists as a variant in French, it may be more salient and thus help identification. Strikingly, in the noisy condition, the native non-aspirated plosives /p/ and /t/ were less easily identified than their aspirated non-native counterparts /pʰ/ and /tʰ/. This might be explained by the shorter voice-onset-time of the native plosives (mean = 34.7 ms) compared to the non-native aspirated ones (110.8 ms), as revealed by acoustic analyses. The native /ʃ/ was also affected by noise, possibly pertaining to its voiceless feature. Preliminary fMRI univariate analyses suggest differential hemispheric lateralization patterns when perceiving native and non-native consonants, in agreement with previous studies. Minagawa-Kawai *et al.* (2005) showed left-lateralized cortical responses for vowel discrimination in native compared to non-native speakers. By contrast, stronger right temporal and sensorimotor cortical activity was reported for non-native/accented than for native phonemes (Wilson & Iacoboni, 2006; Yi *et al.,* 2014). Ongoing multivariate pattern analyses (MVPA) including training classifiers and representational similarity analysis (RSA) will allow to characterize the representational distance of consonants in sensorimotor cortices as a function of the phonological distance between native and non-native phonemes.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., ... & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8, 14.

Arsenault, J. S., & Buchsbaum, B. R. (2016). No evidence of somatotopic place of articulation feature mapping in motor cortex during passive speech perception. *Psychonomic Bulletin & Review*, 23, 1231-1240.

Best, C. T., Tyler, M., Bohn, O., & Munro, M. (2007). Nonnative and second-language speech perception. *Language experience in second language speech learning*, 13-34.

Caballero-Gaudes, C., & Reynolds, R. C. (2017). Methods for cleaning the BOLD fMRI signal. Neuroimage, 154, 128-149.

D'Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex*, 48(7), 882-887.

Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proceedings of the National Academy of Sciences*, 111(19), 7126-7131.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature methods*, 16(1), 111-116.

Minagawa-Kawai, Y., Mori, K., & Sato, Y. (2005). Different brain strategies underlie the categorical perception of foreign and native phonemes. *Journal of cognitive neuroscience*, 17(9), 1376-1385.

Osnes, B., Hugdahl, K., & Specht, K. (2011). Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. *Neuroimage*, 54(3), 2437-2445.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, 103(20), 7865-7870.

Schmitz, J., Bartoli, E., Maffongelli, L., Fadiga, L., Sebastian-Galles, N., & D'Ausilio, A. (2019). Motor cortex compensates for lack of sensory and motor experience during auditory speech perception. *Neuropsychologia*, 128, 290-296.

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and language*, 164, 77-105.

Wilson, S. M., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. *Neuroimage*, 33(1), 316-325

Yi, H. G., Smiljanic, R., & Chandrasekaran, B. (2014). The neural processing of foreign-accented speech and its relationship to listener bias. *Frontiers in human neuroscience*, 8, 768.

A dynamic nasalance analysis of /h̃/ in Zuberoan Basque

Ander Egurtzegi,[1,2] Andrea García-Covelo[5,2,3] & Iñigo Urrestarazu-Porta[1,2,3,4]

[1]*Centre national de la recherche scientifique (CNRS);* [2]*IKER-UMR5478;* [3]*University of Pau (UPPA)*
[4]*University of the Basque Country (UPV/EHU) &* [5]*Institute for Phonetics and Speech Processing (IPS), LMU Munich*
ander.egurtzegi@iker.cnrs.fr,
inigo.urrestarazu-porta@iker.cnrs.fr & andrea.garcia@phonetik.uni-muenchen.de

Introduction. Two endangered Eastern varieties of Basque (basq1248), Zuberoan (Hualde 1993) and Mixean (Camino 2016), show evidence of an extremely rare phonological opposition between an oral /h/ and a nasalized aspirate /h̃/ (Egurtzegi 2018; Egurtzegi 2023). This opposition has only been tentatively proposed for a couple of languages, and some authors assumed it was theoretically impossible (see Walker and Pullum 1999). The impossibility attributed to /h̃/ is rooted in the aerodynamic definition of nasality; i.e. whether or not enough air-stream can go through the nasal cavity once it has produced glottal friction and after it has been divided between the nasal and the oral tract. However, nasalized aspirates can also be theoretically interpreted from an articulatory perspective, where any sound produced with a lowered velum can be considered nasal (Walker and Pullum 1999). Nonetheless, the question remains of whether anything else is required for the functional establishment of a /h/ vs. /h̃/ contrast in a given language. In the Eastern Basque varieties, some aspirates are produced with audible nasalization that spans a whole [VCV] sequence, resulting in sequences that have been impressionistically described as [Ṽh̃iṼ] (Larrasquet 1939), where nasality is phonologically analyzed as originating in /h̃/ and then spreading to the surrounding vowels (Hualde 1993; Egurtzegi 2018). After pointing to similar descriptions of the phonetic realization of /h̃/ such as Madí (jama1261, Amazonas, Brazil) and Yiné (Piro, yine1238, Peru), Blevins and Egurtzegi (2023) proposed that ambient nasalization of the vowels surrounding /h̃/ might be a necessary condition for the stabilization of this segment in a language. However, due to the overall rarity of /h̃/ and the difficulty of obtaining recordings of the languages that show it (most if not all of which are endangered), few studies present phonetic evidence of the realization of /h̃/ (or the /h/ vs. /h̃/ contrast). While Egurtzegi, García-Covelo, and Urrestarazu-Porta (2023) first presented static nasalance-based evidence of this opposition in Zuberoan Basque, here, we present a dynamic analysis of the same dataset. We show that nasalization tends to span a whole VCV sequence when the intervening consonant is /h̃/, with nasalance values that contrast with these in non-nasalized VhV sequences.

Methods. The recordings were made in the Zuberoan village of Larraine, where local participants performed an elicitation task including words containing an aspirate as well as distractor items. We recorded the utterances of 6 volunteer native speakers of Zuberoan Basque (5 male, 1 female; mean age 65, range 60-70) using a Glottal Enterprises nasalance device with a separator handle (NAS-1 SEP), which consists of two microphones separated by a wooden plate that facilitates the separation of the acoustic signal coming from the mouth and that coming from the nose. The stimuli were presented, randomized and recorded with the *SpeechRecorder* software.

Each participant recorded around 102 tokens with an aspirate and another 56 words with no aspirates as fillers. The analysis reported here only involves word-medial intervocalic VHV sequences, where H is one of the two aspirates and V can be any of the 6 vowels in Zuberoan Basque (/a, e, o, i, y, u/). In total, there are 17 tokens with an etymologically oral /h/ (e.g. *soho* 'cropland' from Lat. *solu*(*m*)), 43 with an etymologically nasalized /h̃/ (e.g. *aha̰te* 'duck' from Latin *anate*(*m*)), and 23 including an aspirate and a nasal obstruent in the same word, which has been argued to cause long distance nasal assimilation (e.g. *nahi* 'will' and *ihun* 'nowhere'; Egurtzegi 2018). Crucially, aspirates were all prompted written with no marking for nasalization (i.e., all were written as <h>), irrespective of their etymology and alleged phonological status. Thus, any degree of nasalization in the uttered aspirates should be attributed to the participants' intended pronunciation.

The stereo nasalance data was processed using Praat. For the acoustic analysis, both the nasal and oral channels were band-pass filtered (80 Hz-10000 Hz) and the nasalance (ratio of the nasal amplitude to the sum of the oral and nasal amplitudes, i.e. $A_n/(A_o + A_n) \times 100$) (Carignan 2018) was computed every 5 ms. Curves for the production of each VHV sequence were submitted to a functional principal component analysis (fPCA) (Gubian, Torreira, and Boves 2015). 4 principal components (PCs) cumulatively explained 99% of the variation in the data: PC1 accounted for 90%, PC2 for

7%, PC3 for 2% and PC4 for 0.9%. We only used PC1 for further analysis, given that it explains most of the variation in the data. The variation captured by PC1 is vertical, i.e. it corresponds to the degree of nasality in each production.
A Bayesian multilevel model was fitted in *R* using the *brms* interface to *Stan*. It included the scores of the first PC as response, etymological category as a predictor, and by-speaker correlated varying slope and intercept adjustments and by-word varying intercept adjustments. We used weakly informative priors for the predictor and the intercept, and default priors from *brms* for the random effects and the correlation parameter. We had to adapt the delta to 0.99 to ensure correct convergence (no issues, all $\hat{R} = 1.00$).

Results. Posterior distributions of the model allow to clearly distinguish between oral /VhV/ sequences (median estimate = 0.806 [–0.071, 1.66]) and those that underwent nasal assimilation (median estimate = –0.723 [–1.63, 0.245]). The credible interval of /Vh̃V/ sequences involving an etymological /h̃/ (median estimate = –0.221 [–1.38, 0.98]) covers that of assimilated [Ṽh̃Ṽ] sequences almost completely. Nonetheless, there is considerable uncertainty in the sequences including etymological /h̃/, which results in a partial overlap of its credible interval with that of sequences involving an oral /h/ (Figure 1).

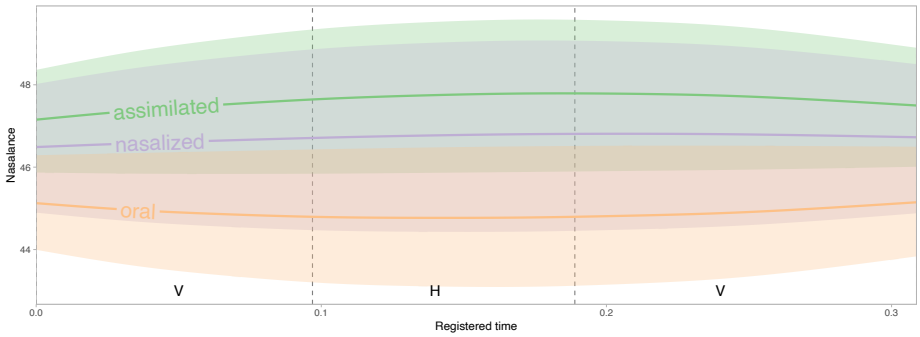


Figure 1: *Nasalance curves of VHV sequences reconstructed with the estimates of PC1 scores. Colored lines indicate the reconstructed curve with the model's median estimate for each category and shades cover 95% credible intervals.*

Discussion. In line with Egurtzegi, García-Covelo, and Urrestarazu-Porta (2023), we suggest that the greater uncertainty of /VHV/ sequences with an etymological /h̃/ and their partial overlap with sequences with oral /h/ is due to two reasons: 1) some speakers have completely lost the /h/ vs. /h̃/ contrast and produce both segments as oral /h/, and 2) a sporadic loss of /h̃/ nasalization in some lexical items among speakers that still maintain the opposition. The results are thus in line with whole /VCV/ sequences being nasalized both in cases of nasal assimilation of /h/ due to a neighboring nasal obstruent as well as in /Vh̃V/ sequences in which /h̃/ is still preserved.

References.

Blevins, J. and A. Egurtzegi (2023). "Refining explanation in Evolutionary Phonology: Macro-typologies and targeted typologies in action". In: *Linguistic Typology* 27 (2), pp. 289–311. DOI: https://doi.org/10.1515/lingty-2021-0036.

Camino, I. (2016). *Amiküze eskualdeko heskuara [The Basque of the region of Amiküze (Mixe)] (Mendaur 11)*. Bilbao: Euskaltzaindia.

Carignan, C. (2018). "Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels". In: *Journal of the Acoustical Society of America* 143 (5), pp. 2588–2601.

Egurtzegi, A. (2018). "On the phonemic status of nasalized /ʰh/ in Modern Zuberoan Basque". In: *Linguistics* 56, pp. 1353–1367.

— (2023). "/ʰh/ hasperen sudurkarituaren inguruan [On the nasalized aspiration /ʰh/]". In: *International Journal of Basque Linguistics and Philology (ASJU)* 57.

Egurtzegi, A., A. García-Covelo, and I. Urrestarazu-Porta (2023). "A nasalance-based study of the /h/ vs. /ʰh/ oposition in Zuberoan Basque". In: *Proc. of the 20th ICPhS*. Ed. by R. Skarnitzl and J. Volín. Prague: Guarant International, pp. 3427–3431.

Gubian, M., F. Torreira, and L. Boves (2015). "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts". In: *Journal of Phonetics* 49, pp. 16–40. DOI: https://doi.org/10.1016/j.wocn.2014.10.001.

Hualde, J.I. (1993). "Topics in Souletin phonology". In: *Generative Studies in Basque Linguistics*. Ed. by J.I. Hualde and J. Ortiz de Urbina. Amsterdam: John Benjamins, pp. 289–327.

Larrasquet, J. (1939). *Le basque de la Basse-Soule orientale*. Paris: C. Klincksieck.

Walker, R. and G.K. Pullum (1999). "Possible and impossible segments". In: *Language* 75 (4), pp. 764–780.

Effects of phonetic contexts on aerodynamic conditions for uvular trills in French

Andres Felipe Lara[1], Didier Demolin[1], Claire Pillot-Loiseau[1]

[1]Laboratoire de Phonétique et Phonologie (CNRS et U. Sorbonne Nouvelle)
[andres.lara, didier.demolin, claire.pillot]@sorbonne-nouvelle.fr

Introduction. In speech, variations in intraoral pressure (Po) result from contextual factors. This includes coarticulation with sounds of varying impedance, adjacent nasals, stress, and speaking rate. Previous research by Lewis (2004) revealed a correlation between the degree of pre-rhotic stricture in consonantal contexts and the likelihood of producing a voiced alveolar trill. However, post-vocalic and absolute-initial contexts did not exhibit the same effects. The present study explores whether similar coarticulatory patterns occur in post-rhotic vocalic contexts with varying degrees of stricture. This study investigates the impact of vowels and syllabic position on the aerodynamic conditions necessary for French uvular trill production. Aerodynamically, trill production requires maintaining a threshold between intraoral pressure (Po) and atmospheric pressure (Pa) for at least 70 ms. Studies by Solé (2002) and Demolin & Van de Velde (ms) demonstrated that trilling in alveolar trills is extinguished when intraoral pressure falls below a threshold of approximately 2.5 hectopascals (hPa). Uvular trills require sustaining a threshold above 2 hPa, with observed thresholds reaching 3.2 hPa (Demolin & Van de Velde, ms). Additionally, trills are characterized by a series of oscillations, with alveolar trills having 2 to 8 and uvular trills having 2 to 3, depending on context and language, Demolin & Van de Velde (ms).

Methods. Aerodynamic data from the "Speech aerodynamic database" (Demolin et al., 2019) was used for this study. Intraoral pressure (Po) was measured using the Physiologia workstation for simultaneous acquisition (Teston and Galindo, 1995). Three French native speaker, comprising of two males and one female were recruited to produce five repetitions of the logatomes "rara", "ruru," and "riri". The recordings were annotated in Praat, incorporating annotations for both the absolute-initial and intervocalic positions. Subsequently, a script was applied to capture measurements at 101 steps along the entire segment. The segment's duration of the curve above a 2 hPa threshold were measured for all tokens as indicators of ideal conditions for trilling. Beats observed were then manually counted after the initial rise and dip of intraoral pressure (Po). Productions with two beats or more were assigned a Trill value, anything under the 2 hPa threshold was deemed an approximant, and anything above the threshold with 1 beat or less was considered a fricative. Further predictions were made using a Bayesian model fitted for categorical regression, a function from Bambi's sub-package *interpret* (Capretto, 2020), incorporating a categorical value (mode of articulation) and a continuous variable (time above the 2 hPa threshold); 4 chains for 1000 tune and 1000 draw iterations (8000 draws total).

Results. See figure 1 for individual productions. In the initial position, Speaker 1 predominantly produces fricatives and trills for the context "rara," with only one approximant. For "riri," the speaker produces three approximants and two fricatives. In the case of "ruru," fricatives are predominantly produced, along with one approximant and one trill. Speaker 2 mainly produces approximants for "rara," with one trill and one fricative. For "riri," only fricatives are produced. In the case of "ruru," predominantly fricatives are produced, with two trills also present. Speaker 3, the only female participant in the study, demonstrates a tendency to produce trills. She exclusively produces trills for "rara." For "riri," one trill and one approximant are produced, while the remaining productions are fricatives. As for "ruru," only trills are produced. In intervocalic position, Speaker 1 tends to produce trills. The speaker produces fricatives and trills for "rara," and exclusively produces trills for "riri" and "ruru." On the other hand, Speaker 2 tends to produce fricatives in intervocalic positions across all contexts, with the exception of one trill for "riri" and "rara." For "ruru," Speaker 2 predominantly produces fricatives, with two trills also present. Speaker 3, similar to Speaker 1, produces three trills and two fricatives for "rara." The only instance where the speaker produces approximants is for the "riri" context, with the remaining sounds being predominantly fricatives, along with one trill. Speaker 3 exclusively produces trills for "ruru." Figure 2 demonstrates the posterior probabilities for the model. The predictions indicate that the vocalic context "rara" is the most favorable for producing trilling; followed by "ruru", and then "riri" which shows very little probability of trilling. The probability of trilling is highest when the Po is sustained above 2 hPa for a longer period of time in the case of "rara" and "ruru". In the case of "riri," there is a conspicuous inclination towards the production of fricatives beyond the 70 ms limit, while the production of trills beyond the 130 ms limit shows an equal likelihood. The words "rara" and "riri" demonstrate highly favorable conditions for the occurrence of approximants under the 50 ms limit whereas "ruru" necessitates an even lower threshold of 30 ms. When comparing positions, it becomes evident that there is a greater likelihood of producing

trill in the initial position as opposed to the intervocalic position. Additionally, the initial position demonstrates a more distinct probability for the production of approximants below the 50 ms limit.

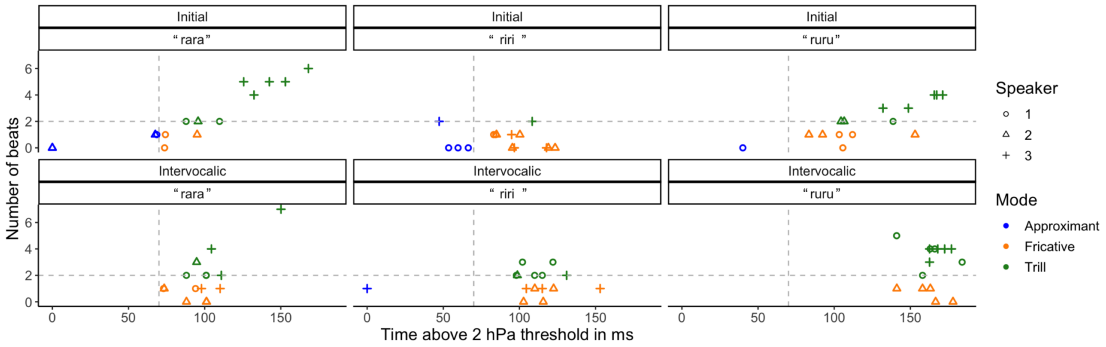


Figure 1: *Rhotic productions in initial and intervocalic contexts (all tokens included). Productions are categorized by production mode (colors) and speaker (shapes). x-axis: duration (in ms) the Po is maintained above the 2 hPa threshold, with a reference line at 70ms. y-axis: number of oscillations achieved.*

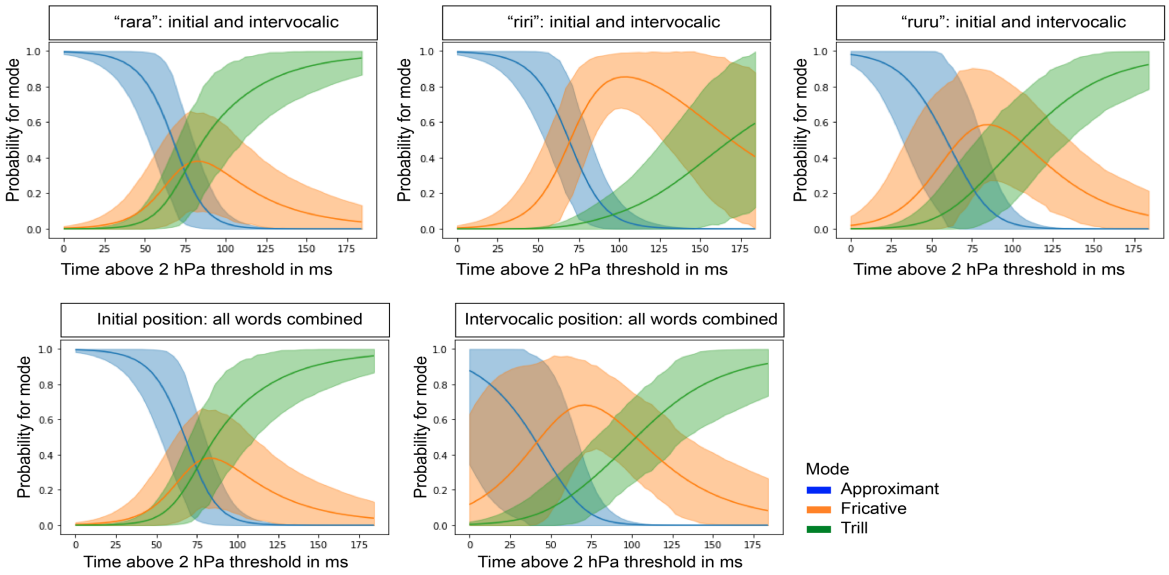


Figure 2: *Categorical regression showing posteriors for the mode of production of rhotics focusing on vocalic context (top tier) and the distinction between initial and intervocalic positions (lower tier). y-axis: probability of each mode;1 indicates highest probability relative to Po being sustained above 2 hPa: x-axis.*

Discussion. The French rhotic exhibits extensive allophonic variation, which includes trill, fricative, or approximant. This variation is influenced by vocalic context, syllabic position, individual preferences, and articulatory configurations. A sustained trill requires the tongue and uvula to assume the correct shape, position, and compliance, along with sufficient oro-pharyngeal pressure building behind the stricture. Coarticulation, voicing, position, and duration can help predict specific allophones. The context [ʀaʀa] is more conducive to successful trill production in both word-initial and intervocalic position, followed by [ʀuʀu], and finally [ʁiʁi], which favors fricatives over trills. The word-initial position exhibits greater favorability for producing trills compared to the intervocalic position. We also found that analyzing aerodynamic data with a 2 hPa threshold facilitated the identification of successful trilling in our three participants. Nevertheless, it's important to note that even under ideal aerodynamic conditions, there are instances when trilling doesn't happen. Our findings also suggest that longer-sustained hPa thresholds lead to more favorable conditions for trilling in [ʀaʀa] and [ʀuʀu], with a suggested time limit of 110 ms for [ʀaʀa] and 140 ms for [ʀuʀu]. Solé (2002) observed that voiced trills are not as strong as voiceless trills, making voiceless trills easier to sustain but less distinct auditorily. Longer durations, which indicate a speaker's ability to sustain a trill, are associated with voiceless trills and exhibit a greater number of beats (Lewis, 2004; Solé, 2002). Future analyses will incorporate F0 data. This study is significant as it sheds light on the phonetic and phonological complexities of trills, offering insights for categorizing and analyzing rhotic productions in French. It also contributes to articulatory modeling and synthesis.

References

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2020). Bambi: A simple interface for fitting Bayesian linear models in Python. arXiv preprint arXiv:2012.10754.

Demolin & Van de Velde. (in press). The quantal change of alveolar [r] to uvular [ʀ]. Language. Advance online publication.

Demolin, D., Hassid, S., Ponchard, C., Yu, S. and Trouville, R. (2019). Speech aerodynamics database. Laboratoire de phonétique et de phonologie, CNRS-MR 7018, Sorbonne Nouvelle, Paris 3, ILPGA. https://corpus.ilpga.fr/aerodynamics

Lewis, A. M. (2004). Coarticulatory effects on Spanish trill production. In Proceedings of the 2003 Texas Linguistics Society Conference (Vol. 116, p. 127). Somerville, MA: Cascadilla Proceedings Project.

Solé, M. J. (2002). Aerodynamic characteristics of trills and phonological patterning. Journal of phonetics, 30(4), 655-688.

Teston, B., & Galindo, B. (1995). A diagnostic and rehabilitation aid workstation for speech and voice pathologies. In Fourth European Conference on Speech Communication and Technology.

Relating frication to articulation in Standard Mandarin apical vowels

Sean Foley[1], Bowei Shao[2], Matthew Faytak [3]

[1] *University of Southern California*
[2] *École Normale Supérieure-Université PSL*
[3] *University at Buffalo*
seanfole@usc.edu, bowei.shao@ens.psl.eu, faytak@buffalo.edu

Introduction. Sibilants are sounds characterized by the production of audible turbulent airflow. Mechanical models of sibilants dictate that the production of turbulent airflow requires both the formation of a narrow constriction in the vocal tract and air projected at a certain velocity through this constriction (Catford 1977; Shadle 1990). These aerodynamic principles suggest that in connected speech the production of frication noise rests on a certain balance being struck between these two factors, i.e. a larger constriction necessitates greater volume velocity and vice versa. We investigate this relationship between lingual constrictions and aerodynamics in Standard Mandarin apical vowels using both articulatory and acoustic data. Previous research has shown that both apical vowels have a lingual posture that closely resembles that of the preceding sibilant with which they are homorganic (Faytak and Lin 2015; Lee-Kim 2014). In addition, there is some debate on whether or not the segments have frication noise targets, though this has only been analyzed impressionistically. To further explore the mechanics of the SM apical vowels, we looked at sequences where each segment occurs adjacent to the sibilant they are homorganic with on *both* sides. Given previous research, there are a number of potential hypotheses of what would occur in such sequences. If both apical vowels have frication noise targets, we would likely see no lingual adjustment as well as little to no change in frication noise during the entire sequence. If both segments lack frication noise targets, we should see a sizeable drop in frication during the apical vowels, comparable to that of other vowels. The general expectation is that such a drop should be accompanied by an increase in the channel size, i.e. tongue tip lowering, though a non-lingual adjustment is also possible, e.g. manipulation of the volume velocity or cavity expansion.

Methods. Simultaneous ultrasound tongue imaging and audio data were collected from five native speakers of Standard Mandarin. The stimuli used in the study consist of disyllabic nonce words. The target segments in the first syllable are the four vocalic segments [ɿ ʅ i a u], with three different onsets [s ʂ ɕ]. The second syllable consists of the consonants [s ʂ ɕ] with [a] serving as the nucleus. Target sequences are those containing the apical vowels flanked on both sides by a homorganic sibilant, i.e. [sɿ.sa] and [ʂʅ.ʂa], with other sequences containing [i a u] in the first syllable used for comparison. For the articulatory data, smoothing-spline ANOVAs (SSANOVAs) were generated comparing the midpoints of the first homorganic fricative (C_1), apical vowel, second homorganic fricative (C_2), and final [a]. Additionally, to quantify constriction degree (CD), the distance between the tongue front registration line data, which tracks the relative tongue front position over time, and points on the palate traced by speaker was calculated. Separate points were selected for alveolar and post-alveolar constrictions, and all values were z-scored across speakers. Generalized Additive Mixed Models (GAMMs) were fit on this data to model temporal changes in CD for the target sequences. For the acoustic data, zero-crossing rate (ZCR) was used as a measure of frication. ZCR measures the number of crossings of zero dB per second in the waveform without relying on voicing or pitch, and has been used to gauge frication levels in similar segments. GAMMs were constructed to model the dynamics of ZCR in the target sequences using values z-scored across speakers. We constructed a single model to model all sequences and report the estimated differences. In sum, there are two articulatory measures, SSANOVAs, which capture overall lingual posture at the segment midpoint, and CD GAMMs, which capture the temporal dynamics of CD. For acoustics, there is a single measure of frication, ZCR, also modeled using GAMMs.

Results. The combined ZCR and CD GAMMs are shown in Figure 1 (right) for four target phrases. Constants were added to separate the two sets of GAMMs for visualization. The acoustic results from ZCR are shown in the bottom set of GAMMs. Two peaks in the ZCR trajectories corresponding to the frication targets of the sibilants [s ɕ ʂ] can be seen,

and two valleys corresponding to the nucleic segments, including vowels [i a u] and apical vowels [ɿ ʅ] are evident. The ZCR values in V_1 position are consistently much lower compared to the two flanking peaks, suggesting that each V_1 has a much lower aperiodic component in the acoustic signal, i.e., frication noise, compared to [s ɕ ʂ]. Significant differences in the ZCR trajectories during the C_1 to V_1 transition and during C_2 are likely attributable to intrinsic differences in sibilant volume velocity and homorganic sequences being more conducive to frication, namely in the [s] to [ɿ] transition. The CD GAMMs show that there is little to no perceptible change in CD during the transition from initial sibilants to the apical vowels, with an increase in channel size occurring during the C_2 sibilant in anticipation of [a]. This is also seen in vowel [i], while during [u] there is a clear sudden increase in channel size. Additionally, the SSANOVAs in Figure 1 (left) show that for the apical vowel target phrases, the tongue blade does not visibly differ in position between the first onset fricative, the apical vowel, and the second onset fricative. Some slight variation in tongue dorsum and blade position between the apical vowel and the second onset consonant can be attributed to anticipation of the upcoming low vowel [a].

Discussion.

To our knowledge, this study presents the first analysis of time-aligned CD and frication measures in apical vowel sequences, highlighting the complex interplay between constriction, frication, and aerodynamics in such sequences. Two major findings are evident in the results. First, during the target phrases, a considerable drop in frication occurs during the apical vowels in V_1 position, reaching the same plateau as the other vowels examined. This result is highly suggestive of both apical vowels lacking frication noise targets. Second, virtually no change in CD occurs during the apical vowels in the target sequences, as confirmed by the CD GAMMs and examination of tongue posture at segment midpoints using SSANOVAs. Interestingly, this same result occurred for the vowel [i]. These findings are surprising, starting from the expectation that such a drop in frication should be due to some lingual adjustment, perhaps with the intention to increase channel size. However, it is possible that during the target sequences, speakers may instead use some non-lingual adjustment to suppress frication so as not to significantly interrupt the current arrangement of the articulators in anticipation of the following sibilant. Sibilants are known for requiring a precise arrangement of the articulators, with constraints put on both the tongue body and tongue front. One potential hypothesis is that speakers are directly manipulating the rate of airflow in the vocal tract during the apical vowels. This would indicate the presence of airflow velocity targets separate from constriction degree targets. However, it is also possible that the initiation of voicing during the apical vowels disrupts the airflow velocity. Incorporating the voiced fricative [ʐ] before the apical vowel [ɻ] into the stimuli would allow for testing this hypothesis. If the trajectory of frication during these sequences does not differ from those observed here, that would suggest other mechanisms are at play here.

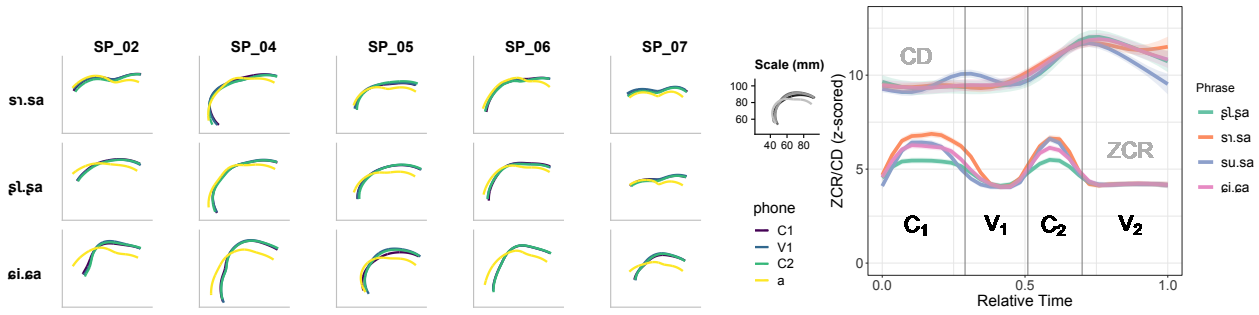


Figure 1: (Left) Tongue surface SSANOVA splines for segment midpoints in homorganic target sequences. Anterior is right in each figure. (Right) Combined ZCR and CD GAMMs fit across speakers for four target sequences.

References.

Catford, John (1977). *Fundamental problems in phonetics*. Midland Books.

Faytak, Matthew and Susan Lin (2015). "Articulatory variability and fricative noise in apical vowels." In: *ICPhS*.

Lee-Kim, Sang-Im (2014). "Revisiting Mandarin 'apical vowels': An articulatory and acoustic study". In: *Journal of the International Phonetic Association* 44.3, pp. 261–282.

Shadle, Christine H (1990). "Articulatory-acoustic relationships in fricative consonants". In: *Speech production and speech modelling* 55, pp. 187–209.

Parametric Excitation of Vocal Tract Resonances by Vocal Fold Motion: A Source-Excitation-Filter Model of Speech Production

Gordon Ramsay[1]

[1]*Spoken Communication Laboratory, Department of Pediatrics, Emory University, Atlanta, Georgia, USA*
gordon.ramsay@emory.edu

Introduction.
The source-filter model of speech production has been highly successful in representing speech as the output of time-varying vocal tract resonances excited by glottal and noise sources, which can be recovered from the speech signal through formant tracking and inverse filtering. Conversely, modern aeroacoustics has provided justification for this model by showing that the Navier-Stokes equations governing fluid flow can be written exactly as a convected wave operator excited by monopole, dipole, and quadrupole sources generated by volume displacement, boundary forces, and shear stresses. However, aeroacoustic models typically assume that source regions are compact and wave operators are quasi-stationary, whereas sources of sound during speech are distributed in space and buried deep within a resonator that changes rapidly in time with articulator and vocal fold motion. Aeroacoustic simulations fail to match real speech, and it is still not clear exactly how motion of the vocal folds and aerodynamic flow through the glottis are converted into sound, which is often circumvented in simulations by assuming heuristically that the source resembles the derivative of the glottal flow imposed as a boundary condition. In this study, we demonstrate that the resonances of the vocal tract are driven not only by external aeroacoustic sources but also by *parametric excitation* that arises from rapid modulation of the vocal tract eigenmodes by motion of the vocal folds, suggesting a new source-excitation-filter model of speech production.

Methods.
The vocal tract can be modelled as a time-varying tube of length L evolving over time T, described on a bounded domain $\Omega = \{(x,t) : x \in [0,L], t \in [0,T]\}$ in $\mathbb{R}^2$, where x is the distance along the mid-line from lungs to lips, and t is time. Assume that the tube shape is given by a cross-sectional area function $A : \Omega \rightarrow \mathbb{R}_+$, and that the physical state of the air in the tube can be represented by functions $\rho, P, U : \Omega \rightarrow \mathbb{R}$, describing the fluid density, pressure, and particle velocity averaged over the cross-section. The global flow field consists of compressible perturbations $\rho_1, P_1, U_1 : \Omega \rightarrow \mathbb{R}$, describing the sound field, superimposed on an underlying mean flow $\rho_0, P_0, U_0 : \Omega \rightarrow \mathbb{R}$, describing respiratory flow, and the sound field is assumed to be driven by aeroacoustic source fields $Q_u, Q_p : \Omega \rightarrow \mathbb{R}$, describing the rate at which mass and momentum are injected per unit length by coupling with the mean flow. Acoustic variables can be expressed as dimensionless groups $p, u, q_u, q_p, \alpha : \omega \rightarrow \mathbb{R}$, defined on a domain $\omega = \{(\xi,\tau) : \xi \in [0,1], \tau \in [0, cT/L]\}$, where c is the sound speed, $\xi = x/L$, $\tau = ct/L$, $\alpha = A/L^2$, and $p = P_1/\rho_0 c^2$, $u = U_1/c$, $q_p = Q_p/\rho_0 c^2$, $q_u = Q_u/\rho_0 c$. Neglecting convective effects, the conservation laws defining quasi-1D mass and momentum balance for the acoustic field are then:

$$\frac{\partial}{\partial \tau}\alpha p + \frac{\partial}{\partial \xi}\alpha u = q_u, \tag{1}$$

$$\alpha\frac{\partial u}{\partial \tau} + \alpha\frac{\partial p}{\partial \xi} + \kappa_r u = q_p, \tag{2}$$

where κ_r is a loss coefficient describing viscous losses per unit length. As boundary conditions, assume that $p(\xi,0) = u(\xi,0) = 0$; that the entrance to the lungs is closed, with $u(0,\tau) = 0$; and that the termination at the lips can be modelled by the usual radiation impedance linking $p(1,\tau)$, $u(1,\tau)$. Equations (1)-(2) describe a linear time-varying system that can be rewritten in operator form with state and source functions $z = (p,u)$ and $q = (q_u, q_p)$ as:

$$\dot{z}_t = \mathcal{L}_t z_t + q_t. \tag{3}$$

The behaviour of the acoustic system is determined by the spectrum of the operator $\mathcal{L}$, given by eigenvalues $\{\lambda_t^i\}$ and eigenfunctions $\{\phi_t^i\}$, and its adjoint $\mathcal{L}_t^\star$, with eigenvalues $\{\overline{\lambda_t^i}\}$ and eigenfunctions $\{\psi_t^i\}$. The eigenfunctions of $\mathcal{L}_t$ and

$\mathcal{L}_t^\star$ form a complete biorthogonal basis for the state space, with $\langle \psi_t^i, \phi_t^j \rangle = \delta_{ij}$. Denoting by e_t^i the projection of z_t onto the invariant subspace of $\mathcal{L}_t$ associated with λ_t^i, equation (3) can be diagonalized as the equivalent system:

$$e_t^i = \langle \psi_t^i, z_t \rangle, \tag{4}$$

$$\dot{e}_t^i = \lambda_t^i e_t^i + \langle \psi_t^i, q_t \rangle - \textstyle\sum_{j \neq i} \langle \psi_t^i, \dot{\phi}_t^j \rangle e_t^j, \tag{5}$$

$$z_t = \textstyle\sum_i e_t^i \phi_t^i. \tag{6}$$

The original infinite-dimensional system of partial differential equations describing the spatio-temporal evolution of the entire acoustic field can thus be reduced to a countable collection of ordinary differential equations describing the temporal evolution of the modal projections. The eigenvalues λ_t^i are the time-varying complex poles of the vocal tract, and define the instantaneous formant frequencies and bandwidths. The eigenfunctions ϕ_t^i represent the instantaneous spatial distribution of pressure and flow rate associated with each formant, whereas the adjoint eigenfunctions ψ_t^i determine the proportion of energy entering each eigenmode from a spatial distribution of mass and momentum sources. The eigenvalues occur in complex conjugate pairs, and each eigenmode behaves as a simple harmonic oscillator, driven by a source term injecting energy from the aerodynamic flow, as well as an additional parametric excitation term that depends on the rate at which the eigenfunctions change in time, which disappears when the operator is self-adjoint or time-invariant. This is mathematically equivalent to a continuous-time parallel formant synthesizer with parameters derived automatically from the area function, except for the presence of the parametric excitation term in (5), which has previously been ignored.

Results.
A finite-volume simulation was constructed to explore the source and excitation terms by integrating equation (3) over a discrete grid $\{(\xi_j, \tau_k) : j = 1 \dots M, k = 1 \dots N\}$. The resulting implicit matrix recursion was solved numerically to synthesize the sound field and speech signal for static vowels and short utterances, using Mermelstein's model to generate a 3D vocal tract shape and Titze's model to generate 3D vocal fold motion. By diagonalizing the system matrices at each point in time to derive discrete-time analogs of equations (4)-(6), the time-varying eigenvalues and eigenfunctions were calculated, along with the modal source, excitation, and amplitude waveforms for each individual formant. Figure 1 shows an example of the modal decomposition for a single formant F2 of the vowel /a/ over several glottal cycles. As the vocal folds open and close, the formant frequency and bandwidth vary periodically; the true aeroacoustic source term follows the left eigenfunction, whereas the parametric excitation term is modulated by the right eigenfunction derivative. The source and excitation terms together excite the eigenmode to generate the formant oscillation, and bear a remarkable resemblance to traditional models of the glottal source waveform, but were derived from first principles from the time-varying area function alone, rather than being imposed as an external boundary condition or pre-determined waveform.

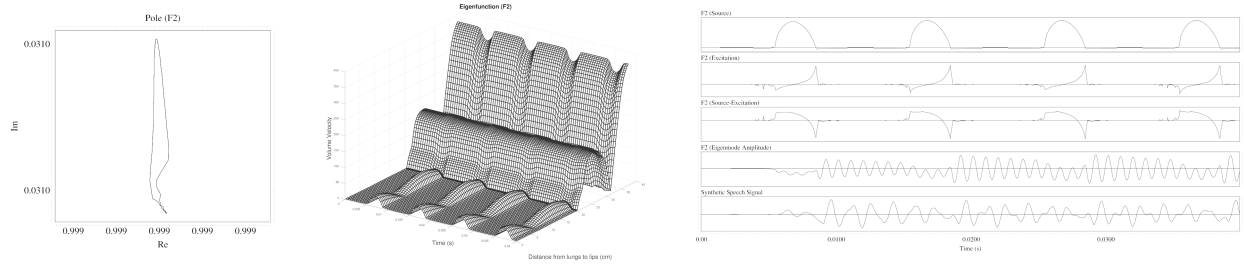


Figure 1: *Simulation results for vowel /a/ showing time-varying eigenvalue, eigenfunction, modal source, parametric excitation, combined modal source and excitation, and modal amplitude for F2, with speech signal at lips.*

Discussion.
Our analysis shows that the conservation laws for a time-varying vocal tract can be rewritten as an equivalent parallel formant synthesizer driven by source and excitation terms that arise respectively from aeroacoustic sources and from parametric excitation, essentially due to the work done by the vocal folds in rapidly modulating the eigenmodes of the acoustic system, consistent with parametric excitation mechanisms found elsewhere in physics. Simulation results indicate that the main impulses exciting each formant at the moments of glottal closure and opening are due to the parametric excitation, not the aeroacoustic source, although this certainly provides energy to the system. We suggest extending the original source-filter model, where sources are specified explicitly as boundary conditions, to a source-excitation-filter model where source and excitation arise implicitly from the dynamics of the vocal tract and the underlying flow field.

Effect of following vowel context on Sevillian derived stop-h sequences

Madeline Gilbert[1]

[1]*Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne Nouvelle*
madeline-besse.gilbert@sorbonne-nouvelle.fr

Introduction. Studies investigating phonetic differences between segments and clusters—-largely comparing duration—-find inconsistent correlations between phonetics and representation. An ongoing sound change in Sevillian Spanish provides a new area in which to investigate this question. In Sevillian, coda /s/ debuccalizes and metathesizes with a following voiceless stop /p, t, k/ (/pasta/: [pahta] → [patha]; Torreira 2006). These derived stop-h sequences [ph, th, kh] look like aspirated stops on the surface, but are representationally different (clusters). While coarse measurements of duration might not distinguish segments from clusters, other measures could reflect hypothesized differences in gestural organization (Browman and Goldstein 1986). More specifically, differences in VOT duration and spectral shape of the release could reflect differences in laryngeal alignment and variability, as well as coarticulation with the following vowel.
This study presents a detailed acoustic analysis of Sevillian stop-h sequences, focusing on the effects of POA and following vowel quality on the duration and spectral shape of the release. Cross-linguistically, VOT duration varies systematically by consonant POA (Cho and Ladefoged 1999) and following vowel quality (Ohala 1983), as does its spectral shape (Suomi 1985; Chodroff and Wilson 2014). These effects are similar across languages, presumably because they derive largely from articulatory and aerodynamic properties inherent to speech production. In particular, these effects reflect the alignment of the glottal opening gesture in relation to the stop closure gesture for aspirated stops. If the effects differ in Sevillian stop-h, this suggests different gestural organization. This Sevillian data provides the basis for future large-scale, cross-linguistic comparisons.

Methods.
24 female Sevillians read sentences containing /C/ and /sC/ target words, matched for phonological context and stress (l**ati**na-l**asti**ma; 1213 word tokens). The duration of VOT was measured (release burst to voicing onset) and is modeled with linear mixed-effects models and *emmeans*. Spectral moments (COG, SD, skew, kurtosis) were taken at 10 equidistant points across the VOT interval using *Praatsauce*, and their trajectories will be modeled with GAMs.

Results.
In this abstract, I describe results for words with following /a, i/, the most articulatorily and aerodynamically distinct vowels. I highlight the principle points here and will present more detailed analyses at the conference.
VOT duration: VOT is longer in /sC/ words than /C/ words, as expected given metathesis (/sp/ = 40ms, /p/ = 15ms; /st/ = 62ms, /t/ = 15ms; /sk/ = 57ms, /k/ = 21ms). For /sC/ words, this means that the observed VOT duration cline is as follows: /st/ (62ms) > /sk/ (57ms) > /sp/ (40ms). The effect of vowel quality is such that VOT is longer before /i/ than /a/, but significantly so only for /st/ (/sti/ = 71ms; /sta/ = 62ms).
Spectral (**Figure 1**): Following vowel quality has little effect on the spectral properties of /st, sp/ releases. Notably, for /st/, COG is high and skew is low, which reflects the ongoing [th] → [ts] affrication change (Ruch 2008). /sk/ shows larger contextual vowel effects. In comparison to /ska/ ([kha]), the release of /ski/ ([khi]) has higher COG and SD and lower skew (higher frequency energy concentration, more diffuse spectrum). The effect of vowel on kurtosis is unclear.

Discussion. The observed patterns are both similar to and different from observed cross-linguistic tendencies. Cross-linguistically, VOT is longer for velars (vs. labials/alveolars) and before high/front (vs. low) vowels. Longer VOT for velars can be attributed to factors like backness of articulation, amount of contact, and articulatory speed (Cho and Ladefoged 1999), which affect the time it takes for oral pressure to drop, creating the transglottal pressure differential necessary for voicing. Longer VOT before high/front vowels is attributed to their narrow release, which also affects how long it takes oral pressure to drop (Ohala 1983). Sevillian diverges from the near-universal VOT cline by POA in that

the release of [th]/[ts] is longer than that of [kh]. This is likely because /st/ is most advanced in the metathesis change. Sevillian partially fits with cross-linguistic patterns for the effect of following vowel, in that VOT is longer before /i/ than /a/ for /st/. However, Sevillian also diverges from the common effect of vowel quality because /i/ does not affect /sk, sp/ in the same way as /st/, and the other vowels have inconsistent effects across consonant POA. That VOT duration is not cleanly attributable to articulation or aerodynamics suggests that other factors—such as low gestural overlap between the glottal and stop closure gestures—may play a role. Low overlap could lead to long VOT and little effect of the following vowel. If VOT is long enough, the drop in oral air pressure necessary for voicing could be achieved before voicing would start, regardless of the quality of the following vowel (suggested for Danish by Mortensen and Tøndering 2013).
For spectral shape, the patterns for /sp, sk/ are similar to findings on aspirated stops in other languages where /sp/ shows little contextual variation and /sk/ shows substantial variation (e.g., Finnish; Suomi 1985). For /st/, it is notable that the spectral shape of the release is similar before all vowels. This suggests that the [th] → [ts] affrication change has spread to all contexts, even though it presumably originated in high-vowel contexts. In comparing the Sevillian results to Greek (Nicolaidis et al. 2019), English (Chodroff and Wilson 2014), and Danish (Puggaard-Rode 2022), I will discuss possible differences in gestural organization between Sevillian stop-h sequences and (un)aspirated stops.

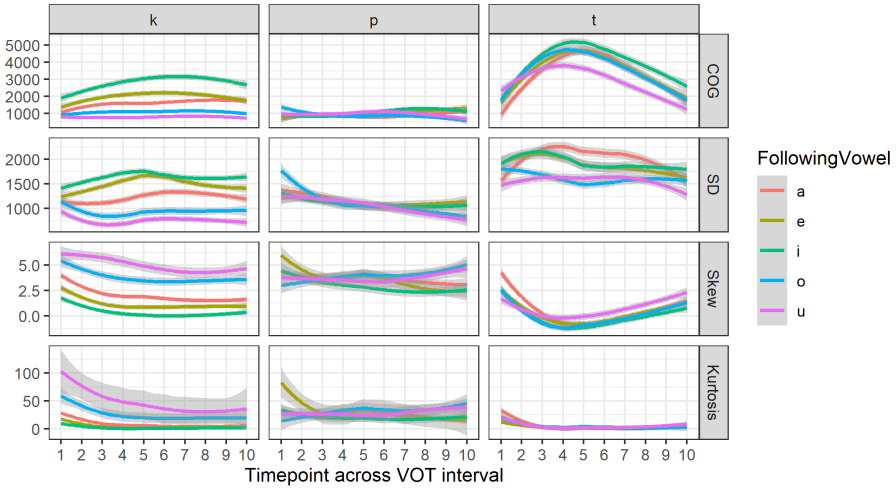


Figure 1: *Spectral properties of Sevillian metathesized /sp, st, sk/ sequences by following vowel.*

References.

Browman, Catherine and Louis Goldstein (1986). "Towards an Articulatory Phonology". In: *Phonology Yearbook* 3, pp. 219–252.

Cho, Taehong and Peter Ladefoged (1999). "Variation and universals in VOT: Evidence from 18 languages". In: *Journal of Phonetics* 27, pp. 207–229. DOI: 10.1006/jpho.1999.0094.

Chodroff, Eleanor and Colin Wilson (2014). "Burst spectrum as a cue for the stop voicing contrast in American English". In: *The Journal of the Acoustical Society of America* 136.5, pp. 2762–2772. DOI: 10.1121/1.4896470.

Mortensen, Johannes and John Tøndering (2013). "The effect of vowel height on voice onset time in stop consonants in CV sequences in spontaneous Danish". In: *Proceedings of Fonetik 2013. The XXVIth Annual Phonetics Meeting*. Ed. by Robert Eklund. Linköping University, Linköping, Sweden, pp. 49–52.

Nicolaidis, Katerina, Anna Sfakianaki, George Vlhavas, and George Kafentzis (2019). "An acoustic study of Greek voiceless stops". In: *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*. Ed. by Sasha Calhoun, Paola Escudero, Marija Tabain, and Paul Warren. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Ohala, John J. (1983). "The origin of sound patterns in vocal tract constraints". In: *The Production of Speech*. Ed. by Peter F. MacNeilage. New York, NY: Springer New York, pp. 189–216. DOI: 10.1007/978-1-4613-8202-7_9.

Puggaard-Rode, Rasmus (2022). "Analyzing time-varying spectral characteristics of speech with function-on-scalar regression". In: *Journal of Phonetics* 95, p. 101191. DOI: 10.1016/j.wocn.2022.101191.

Ruch, Hanna (2008). "La variante [ts] en el español de la ciudad de Sevilla: Aspectos fonético-fonológicos y sociolingüísticos de un sonido innovador". MA thesis. Zürich, Germany: University of Zürich.

Suomi, Kari (1985). "The vowel-dependence of gross spectral cues to place of articulation of stop consonants in CV syllables". In: *Journal of Phonetics* 13.3, pp. 267–285. DOI: 10.1016/S0095-4470(19)30759-4.

Torreira, Francisco (2006). "Coarticulation between aspirated-s and voiceless stops in Spanish: An interdialectal comparison". In: *Selected proceedings of the 9th Hispanic Linguistics Symposium*. Ed. by Nuria Sagarra and Almeida Jacqueline Toribio. Somerville, MA: Cascadilla Proceedings Project, pp. 113–120.

Sex-Specific Patterns in Intraoral Pressure in the Production of Georgian and German Ejectives

Nato Sulaberidze, Adrian P. Simpson

Friedrich Schiller University Jena

nato.sulaberidze@uni-jena.de, adrian.simpson@uni-jena.de

Introduction. In the production of ejective stops, an upward movement of the larynx is commonly proposed to explain the intraoral pressure required for ejective release, as the other articulators, including the glottis and velum, remain closed during the production of these sounds (Catford, 1988; Maddieson, 2013). However, there is evidence that challenges the involvement of the larynx in the production of ejective sounds (Kingston, 1985; Brandt et al., 2021; Sulaberidze et al., 2023a). Georgian contains phonological ejective stops that contrast with voiceless aspirated and voiced stops at labial, dental, and velar places of articulation. Previous findings on Georgian show higher intraoral pressure (IOP) in ejective stops compared to their pulmonic congeners, especially in sentence-initial position (Sulaberidze et al., 2023b). However, there is insufficient evidence that the larynx plays a vital role in the production of Georgian ejectives (Brandt et al., 2021; Sulaberidze et al., 2023a). Although they are not phonological, ejectives are also acoustically and auditorily perceptible in German and English (Brandt et al., 2021; Price et al., 2022; Sulaberidze, et al. 2022). For sociophonetic ejectives in English, there is no consistent evidence of laryngeal raising (Price et al., 2022). In German, larynx elevation is not motivated. Instead, these ejective-sounding final fortis stops are produced epiphenomenally when a word-final fortis plosive is followed by a glottalised onset vowel, as in [hat’ ʔaɐ̯x]. Earlier investigations on German indicated a lower IOP of the fortis stops in target positions (word-final, glottalised) compared to contexts without expected glottalisation (Brandt et al., 2021). Since the precise articulatory and aerodynamic mechanisms of ejective production are still not fully understood, analyzing ejectives from different angles based on their phonological status can provide insights into a comprehensive understanding of the properties involved in their production and, moreover, suggest a possible pathway for their emergence. Our recent study of Georgian, German and English was limited to female speakers. In the present analysis, we explore potential sex-specific differences in IOP in Georgian and German. There are systematic physiological differences in laryngeal and aerodynamic conditions between females and males (Fitch et al., 1999; Kahane, 1978). Sex-specific differences in the temporal characteristics of stops, such as VOT, have also been reported (Robb et al., 2005; Whiteside et al., 2004; Oh, 2011). However, in terms of IOP, no significant differences have been found between female and male speakers in the production of English /p/ (Koenig, 2000).

Methods. In order to measure IOP during the production of stops, a bruxing plate was made for each subject. A thin tube was attached to the plate so that one end was positioned towards the centre of the hard palate but still anterior of the velum. A pressure sensor was attached to the outer end of the tube outside the mouth. Given the position of the tube, measurement is limited to labial and apical stops. In German, we investigated bilabial and alveolar fortis stops in three contexts: (a) word-final before an open onset vowel (target) (e.g.: [hat ʔaɐ̯x]), (b) intervocalic stop followed by /ə/ (control schwa) (e.g.: [hatə ʔaɐ̯x]), and (c) word-final stop followed by a voiced sonorant with expected neutral IOP (control pressure) (e.g.: [hat niː]). For Georgian stops (/b p p’ d t t’/) there were three sentence contexts: sentence-initial (beginning of the utterance), word-initial (onset of the second word in the sentence) and word-medial (nucleus of the second word). Speech of 22 Georgian (9 males and 13 females) and 25 German (11 males and 14 females) subjects were analysed. We used GAMMs (Wieling, 2018) for the statistical analysis of the IOP changes over time (23 measurement points over plosive release at every 10 milliseconds, where plosive release was fixed on the 15th MP).

Results. Separate statistical models for each sound showed the following results: There were no significant sex-specific differences between the IOP curves of Georgian ejectives (/p’/ and /t’/) in any of the sentence contexts (**Figure 1**, last row). Moreover, the curves of none of the sounds in the sentence-initial positions were statistically different between Georgian female and male subjects. However, we observed significant sex-specific differences in /p/ in word-initial and word-medial positions, with the differences in /p/ being due to an earlier onset of the increase in IOP in female speakers **Figure 1** (middle row). Differences in /p/ were also observed in the speech of females and males in German **Figure 1** (top row): German female subjects showed an earlier rise of IOP in /p/ in all contexts and in /t/ only in the control condition (control pressure). There were no significant differences between the curves in German /t/ in other contexts. In terms of IOP peaks, there were no statistically different values in the German or Georgian data, except for the Georgian labial voiced stops (/b/) in word-initial and word-medial positions, where females showed greater peaks than male speakers.

Discussion. We did not expect differences in IOP between females and males. And this expectation was met for the majority of the stops. However, contrary to our expectations, we observed some differences in IOP in /p/ of female and male speakers in both languages, in /t/ in German, and in /b/ in Georgian. The female rise in pressure begins earlier in /p/

and there is a longer plateau before the pressure drop, associated with the stop release. It is possible that in the smaller female supraglottal cavity IOP rises faster and peaks, reaching subglottal pressure earlier, in labial fortis stops (not, however, in Georgian ejectives).

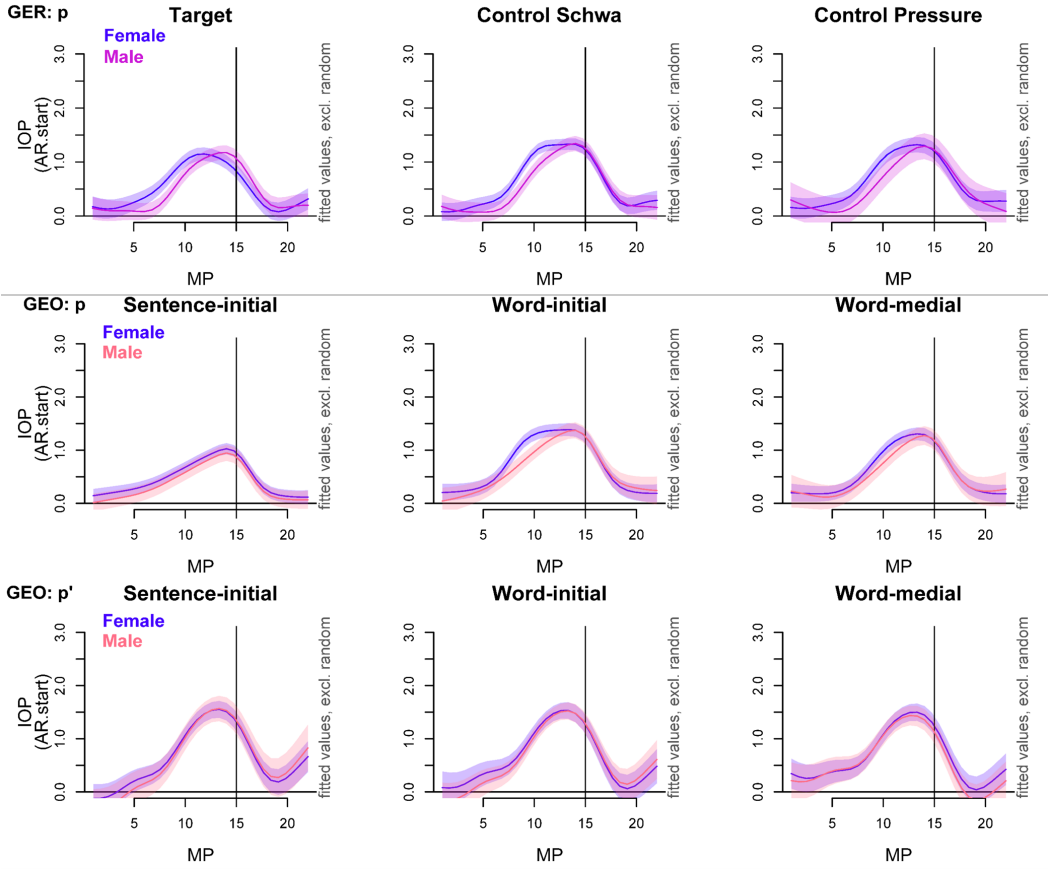


Figure 1: *Sex-specific IOP differences: in German bilabial fortis stop in three sound contexts (top), in Georgian bilabial voiceless (middle) and ejective stops (bottom) in three sentence contexts. Vertical line = stop release.*

References

Brandt, E., & Simpson, A. P. (2021). The production of ejectives in German and Georgian. Journal of Phonetics, 89

Catford, J. (1988). A Practical Introduction to Phonetics. Oxford: Clarendon Press.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: a study using magnetic resonance imaging. *Journal of the Acoustical Society of America*, 106(3), 1511–1522.

Kahane, J. C. (1978). A morphological study of the human prepubertal and pubertal larynx. *American Journal of Anatomy*, 151(1), 11–19.

Kingston, J. (1985). The ineffectiveness of larynx movement. The Journal of the Acoustical Society of America, 77(1), 86–87.

Koenig, L. L. (2000). Laryngeal Factors in Voiceless Consonant Production in Men, Women, and 5-year-olds. Journal of Speech, Language, and Hearing Research, 43(5), 1211–1228.

Maddieson, I. (2013). The World Atlas of Language Structures Online. In M. S. Dryer & M. Haspelmath (Eds.). Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://wals.info/chapter/7

Oh, E. (2009). Effects of speaker gender on voice onset time in Korean stops. *The Journal of the Acoustical Society of America*. 125. 2574.

Price, L., Pouplier, M., & Hoole, P. (2022). Kehlkopfanhebung in englischen wortfinalen Ejektiven: eine Echtzeit-MRT-Studie. P&P 18, Bielefeld

Robb, M., Gilbert, H., Lerman, J. (2005). Influence of gender and environmental setting on voice onset time. *Folia Phoniatrica et Logopedica*, 57(3): 125-33.

Sulaberidze, N., & Simpson, A. P. (2022). Werden deutsche epiphänomenale Ejektive als ‘echte’ Ejektive wahrgenommen? P&P 18, Bielefeld.

Sulaberidze, N., Brandt, E., Hoole, P., Krämer, M., Reichenbach, J. R., & Simpson, A. P. (2023a). Ejectives in Georgian. A real-time MRI analysis of vertical larynx movement. In R. Skarnitzl & J. Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences. Guarant International. 952–956.

Sulaberidze, N., Brandt, Simpson, A. P. (2023b). Intraoral Pressure in Georgian Ejectives. In T. Pistor, C. Steiner, F. Tomascheck, A. Leemann (Eds.), *Book of Abstracts der 19. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, 06.-07.10.2023. Universität Bern: Bern Open Publishing. 63 – 64.

Whiteside, S.P., Henry, L. and Dobbin, R. (2004) Sex differences in voice onset time: a developmental study of phonetic context effects in British English. *The Journal of the Acoustical Society of America*, 116 (2). 1179-1183. ISSN 0001-4966

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, 70, 86–116

Discovering dynamical models of speech using physics-informed machine learning

Sam Kirkham

Lancaster University
s.kirkham@lancaster.ac.uk

Introduction.
Spoken language is characterised by a highly variable and complex set of physical movements that map onto to the small set of discrete units that represent the building blocks of speech. What are the fundamental dynamical principles that underlie this relationship? One approach to this question is Articulatory Phonology/Task Dynamics (AP/TD), which conceptualises speech as series of gestures, with each gesture modelled as a second-order differential equation (Saltzman and Munhall 1989; Browman and Goldstein 1992; Iskarous 2017). Recent studies, however, have critiqued the ability of AP/TD to accurately capture the quantitative facts of empirical data, raising questions around its theoretical validity (Turk and Shattuck-Hufnagel 2020; Elie, Lee, and Turk 2023). Some of these criticisms have been addressed in extensions of AP/TD that model the role of feedback in planning and execution (Tilsen 2016; Parrrell et al. 2019), but in this study we focus solely on the intrinsic dynamics of the gesture once initiated by the nervous system. Our aim is to discover new, accurate and simple task dynamic models of gestures directly from data.

Methods.
Our approach uses sparse symbolic regression (Brunton, Proctor, and Kutz 2016) for discovering the underlying dynamics behind articulatory trajectories. Sparse symbolic regression is a physics-informed machine learning technique that regresses a library of functions (e.g. x, $\dot{x}$, $\ddot{x}$, x^2, x^3, $\cos x$, etc) against empirical data, but uses a sequential thresholding algorithm to filter out non-essential terms. This aims to strike the balance between a good-fitting model and a parsimonious model, thus exposing a small number of important terms that govern the system under study. A key property is that the models also allow for the integration of known physical constraints of the system under study. The model output is a symbolic equation derived directly from empirical data, alongside the parameters that generate a specific trajectory. This allows us to discover new principles of articulatory dynamics underlying speech production and, crucially, examine whether the same dynamics underpin all articulatory movements, or whether this varies between different gestures or even different speakers.
Experiment 1 applies the method to simulated data from known equations, testing whether the models in Saltzman and Munhall (1989) can be recovered from simulated data, even in the presence of variation in (1) trajectory length; (2) initial conditions; (3) different targets; and (4) extensive added noise.
Experiment 2 then extends this approach to real data. Repetitions of the syllable /pa pa pa .../ were extracted from the X-Ray MicroBeam corpus (Westbury 1994). This task was used as it features a high number of repetitions of the same syllable, allowing us to test the effect of (1) small amounts of within-speaker variation; and (2) between-speaker variation. We use the Euclidean distance between the upper and lower lips to calculate lip aperture for /p/. Velocity was calculated as the first derivative of the position data and each gesture was segmented into closure and release gestures at velocity zero-crossings. We use the SINDy implementation of sparse symbolic regression on the segmented position and velocity trajectories to conduct model discovery (de Silva et al. 2020) using a novel ensembling technique that allows us to quantify the distribution of possible models and parameters across a data set.

Results. The results on simulated data show a very high degree of accuracy, recovering the original model in each case and correctly identifying simulated model parameters. See Figure 1 for an example of model performance on normal and noisy simulated data.
Our preliminary results on empirical data suggest a small set of candidate models for speech dynamics. We fit both a first-order model and a second-order model, which allows us to compare the utility of acceleration-based information in model

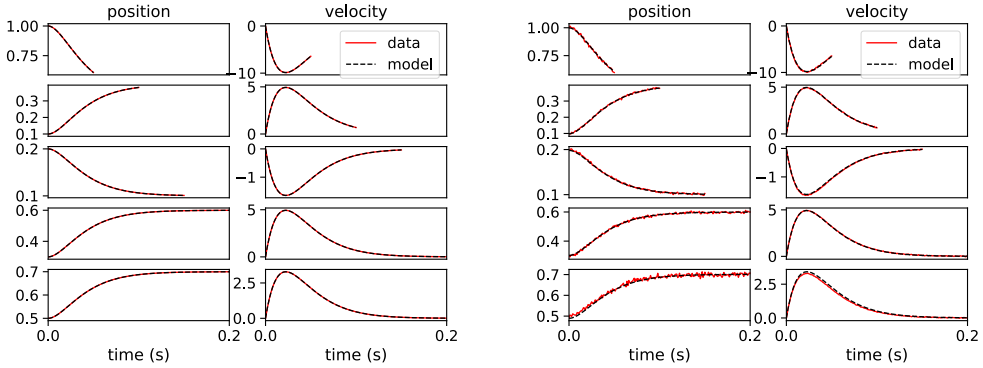


Figure 1: *5 random trajectories showing simulated SM89 data and SINDy model predictions for normal and noisy conditions. Note that the y-axis range varies across each subplot to fit the range of the data.*

fits. In both cases, adding polynomial terms above the quadratic does not result in substantially better reconstructions of the original trajectories.

Discussion.

While the results show that there is more than one good model, different models raise different theoretical implications for how we conceptualise the planning processes involved in gestural execution. For example, the discovered second-order model relaxes the critical damping constraint of Saltzman and Munhall (1989), but it introduces an additional constraint that relates stiffness and damping to gestural duration in a non-linear manner (Shaw and Chen 2019). While this substantially increases the fit between model and data, it introduces additional theoretical constraints that must be considered. For each model, we simulate new data across an exhaustive parameter grid, allowing us to further show which parameter combinations are possible for a given gestural target and duration, with evidence of non-uniqueness in some areas of the parameter space. We conclude by identifying future opportunities and obstacles in data-driven model discovery.

References.

Browman, Catherine P. and Louis Goldstein (1992). "Articulatory phonology: an overview". In: *Phonetica* 49.3-4, pp. 155–180.

Brunton, Steven L., Joshua L. Proctor, and J. Nathan Kutz (2016). "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". In: *Proceedings of the National Academy of Sciences* 113.15, pp. 3932–3937.

de Silva, Brian M., Kathleen Champion, Markus Quade, Jean-Christophe Loiseau, J. Nathan Kutz, and Steven L. Brunton (2020). "PySINDy: A Python package for the sparse identification of nonlinear dynamical systems from data". In: *Journal of Open Source Software* 5.49, p. 2104.

Elie, Benjamin, David N. Lee, and Alice Turk (2023). "Modeling trajectories of human speech articulators using general Tau theory". In: *Speech Communication* 151, pp. 24–38.

Iskarous, Khalil (2017). "The relation between the continuous and the discrete: A note on the first principles of speech dynamics". In: *Journal of Phonetics* 64, pp. 8–20.

Parrrell, Benjamin, Vikram Ramanarayanan, Srikantan Nagarajan, and John Houde (2019). "The FACTS model of speech motor control: Fusing state estimation and task-based control". In: *PLoS Computational Biology* 15.9, pp. 1–26.

Saltzman, Elliot and Kevin G. Munhall (1989). "A dynamical approach to gestural patterning in speech production". In: *Ecological Psychology* 1.4, pp. 333–382.

Shaw, Jason A. and Wei-Rong Chen (2019). "Spatially conditioned speech timing: Evidence and implications". In: *Frontiers in Psychology* 10.2726, pp. 1–17.

Tilsen, Sam (2016). "Selection and coordination: The articulatory basis for the emergence of phonological structure". In: *Journal of Phonetics* 55, pp. 53–77.

Turk, Alice and Stefanie Shattuck-Hufnagel (2020). *Speech Timing: Implications for Theories of Phonology, Phonetics, and Speech Motor Control*. Oxford: Oxford University Press.

Westbury, John R. (1994). *X-Ray Microbeam Speech Production Database User's Handbook*. Madison, WI: Waisman Center.

Spatiotemporal Coupling of the Jaw and Lower Lip: Comparing Talkers with Parkinson's Disease and Amyotrophic Lateral Sclerosis

Mili Kuruvilla-Dugdale[1], Antje S. Mefferd[2]
[1]University of Iowa. IA, USA
[2]Vanderbilt University Medical Center, Nashville, TN, USA
mkuruvilladugdale@uiowa.edu, antje.mefferd@vumc.org

Introduction. Speech motor control is highly complex because the tongue, lips, and jaw move in a carefully timed fashion towards spatial targets within the vocal tract. Within a word, several spatial targets need to be reached in sequence. To allow for a smooth and economical transition from one spatial target to another, articulatory movements of one articulator often overlap with those of another articulator. In previous studies, such inter-articulatory timing patterns have been quantified using phase angles (e.g., Kelso et al., 1995, Nittrouer et al., 1991, Shaiman, 2002). These studies have shown that the degree of gestural overlap is not constant but shifts with task demands (e.g., speech rate, lexical stress).

However, the ability to change the degree of gestural overlap requires capacity to manipulate the strength of inter-articulatory coupling. That is, when inter-articulatory coupling is strong, the articulators move in a more synchronized fashion and gestural overlap is low. However, when inter-articulatory coupling is low, then articulators move more independently and less synchronized.

The flexibility to manipulate coupling strength may be particularly difficult when dealing with impaired speech motor systems. For example, talkers with amyotrophic lateral sclerosis (ALS) are thought to rely more heavily on jaw movements to passively move the tongue towards its spatial target (e.g., Yunusova et al., 2008). A jaw-reliance would suggest that talkers with ALS show relatively low degree of gestural overlap. By contrast, talkers with dysarthria due to Parkinson's disease (PD) may exhibit similar inter-articulatory timing patterns as healthy talkers because they are thought to merely downsize the amplitude of their articulatory movements (e.g., Forrest et al., 1989). So far, however, inter-articulatory timing patterns have rarely been studied in talkers with dysarthria and in remains unclear if and how inter-articulatory timing patterns differ from those of healthy talkers as well as across talkers with different underlying pathophysiologies. Thus, the current study sought to investigate the inter-articulatory timing patterns of the lower lip and jaw in talkers with ALS and PD. Findings will provide new insight in speech motor control characteristics as a first step to better define therapeutic needs.

Methods. So far, kinematic data have been recorded from six individuals with ALS, nine people with PD, and ten controls using a 3D electromagnetic articulography (Wave, NDI Inc.) However, data collection for this project is ongoing and we expect to include more participants to the final dataset. Talkers with ALS and PD ranged in their dysarthria severity from mild to moderate-severe. All participants produced five repetitions of the word "muffin" embedded in the carrier phrase "Say__again". The utterance was chosen because it included a C_1VC_2 sequence that facilitated similar movements of the jaw, lower lip, and tongue tip. To record speech kinematics, small sensors were affixed along the mid-sagittal plane to the articulators (i.e., tongue, jaw, lips). A head reference sensor recorded the head movements. For this study, only the kinematic data of the lower lip and the jaw were used for data analysis. Acoustic data were acquired at a rate of 22 kHz and movement data at the rate of 400Hz. Kinematic data were corrected for head movements and rotated into a head-based coordinate system using software provided by NDI. All kinematic data were low-pass filtered at 15Hz.

Using a custom-written MATLAB script, the vertical movements of the jaw, lower lip, and tongue tip were analyzed. Lower lip and tongue movements were not decoupled from the jaw because this step was not necessary given the purpose of this study and the measurement approach that was taken. Specifically, this study focused exclusively on lag times between the jaw and lower lip as they reached their spatial targets for the open vowel /ʌ/ and the labiodental fricative /f/. Although this approach differs from the traditional phase angle calculations, it is well-suited to quantify the strength of inter-articulatory coupling.

First, the word repetitions were parsed from the carrier phrase. The onset was defined as the positional maxima of the lower lip at the word initial consonant /m/ and the offset was defined as the positional maxima of the tongue tip at the word final consonant /n/. For better spatial alignment, the parsed jaw and lower lip movements were then z-scored and plotted in one graph (see **Figure 1**). Then, an algorithm identified the timepoints of the positional minima for the jaw and lower lip during the vowel /ʌ/ and the timepoints of the positional maxima for the jaw and lower lip during the labiodental fricative /f/. Then, the timepoint of the lower lip was subtracted from the timepoint of the jaw for each target (see **Figure 1**). Finally, all lag times, which consisted of positive and negative values, were converted to absolute numbers because the study sought to determine the strength of lower lip and jaw coupling. The order in which the lip and jaw reached the target was not of interest. Because lag times may be more difficult to interpret when speakers produce the target utterance at different articulatory rates, we also calculated the percent lag time (%lag), which was the lag time relative to the total word duration. To determine the consistency of the lag times across five repetitions, we also calculated the coefficient of variation (CoV) based on the raw lag times (including positive and negative values). The CoV was defined as the standard deviation across five repetitions divided by the speaker's mean lag time across five repetitions.

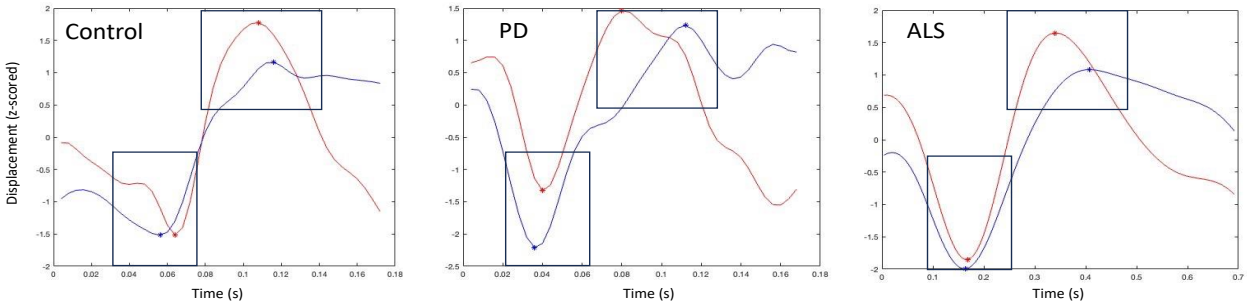


Figure 1: *Spatially normalized (z-scored) movements of the lower lip (red) and jaw (blue) during the word "muffin". Shaded areas point out the timing patterns of the lower lip and jaw for the targets /ʌ/ and /f/.*

Linear mixed models were completed to determine between-group differences in absolute and percent lag times with group (controls, ALS, PD) as the fixed effect, and subject as the random effect. The repeated measures variable consisted of the five repetitions of the word from each participant. For CoV, a between-group ANOVA was used to examine group differences. Because of the preliminary nature of the dataset, the critical alpha-level of $p < .05$ was selected for all tests.

Results. Group means (*SE*) of each dependent variable are provided in **Table 1**. No significant group differences were found for the absolute and the percent lag times as well as for the CoV for either target. However, although the target effect was not tested statistically, group means in **Table 1** show that the lag times for the vowel tended to be shorter than the lag times for the fricative in talkers with PD and controls whereas talkers with ALS had similar lag times for both targets. Finally, as also shown in **Table 1**, the CoV for the vowel tended to be lower in talkers with ALS than in the controls and the talkers with PD.

Table 1: *Group Means (SE) for all dependent variables.*

Group	Lag_/ʌ/	%Lag_/ʌ/	CoV_Lag_/ʌ/	Lag_/f/	%Lag_/f/	CoV_Lag_/f/
Control	.004 (.001)	2.34 (.425)	.739 (.555)	.012 (.002)	6.26 (1.16)	-.780 (.683)
PD	.007 (.002)	3.37 (.509)	2.47 (.663)	.013 (.003)	6.66 (1.39)	.234 (.720)
ALS	.010 (.003)	1.98 (.777)	.263 (.716)	.009 (.004)	2.95 (2.12)	-.935 (.882)

Discussion. The current study sought to determine potential differences in the strength of inter-articulatory coupling between talkers with ALS, PD, and controls. Furthermore, the study investigated the extent to which inter-articulatory coupling was consistent across five repetitions of the same utterance and compared these findings across the three groups. It was hypothesized that talkers with ALS would exhibit stronger inter-articulatory coupling than talkers with PD and controls based on the notion that their articulators are differentially affected by the disease (e.g., Langmore & Lehman, 1994) and therefore, these talkers may rely more on the jaw to move their lower lip and tongue. Preliminary findings for lag times did not support this hypothesis as between-group tests did not reach statistical significance. It is, however, interesting that the talkers with ALS had similar lag times for both targets, which suggest that their ability to alter the strength of inter-articulatory coupling may indeed be constrained. Talkers with ALS also tended to show more consistent lag times than controls and talkers with PD given the trends in the CoV values for the vowel target. This finding also aligns with the idea that talkers with ALS have less flexibility to change the coupling of the lower lip and jaw. However, further research is warranted to support this notion.

Talkers with PD were expected to exhibit similar articulatory coupling behaviors as healthy talkers and the preliminary findings supported this hypothesis. However, for the vowel, lag times tended to be longer in talkers with PD than in controls suggesting that lower lip and jaw may be less tightly coupled. Although such findings have been reported for fast rate effects (e.g., Nittrouer, 1991), word durations did not differ between talkers with PD and controls suggesting that the trend for longer lag times in the PD group for the vowel target is not an epiphenomenon of a faster rate. The CoV values of talkers with PD also tended to be greater than those of controls and talkers with ALS. Thus, talkers with PD may have the flexibility to alter their inter-articulatory coupling strength; however, they may have difficulty controlling these inter-articulatory timing patterns. It should be emphasized, again, that larger sample sizes are needed to solidify the observed trends and re-examine their statistical significance once adquate statistical power.

References

Forrest, K., Weismer, G., & Turner, G. S. (1989). Kinematic, acoustic, and perceptual analyses of connected speech produced by Parkinsonian and normal geriatric adults. *Journal of the Acoustical Society of America*, *85*(6), 2608-2622.

Kelso, J., Vatiskiotis-Bateson, E., Saltzman, E., & Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America, 77,* 266-280.

Nittrouer S. (1991). Phase relations of the jaw and tongue tip movements in the production of VCV utterances. *Journal of the Acoustical Society of America, 90*(4)*,* 1806-1815.

Shaiman, S. (2002). Articulatory control of vowel length for contiguous jaw cycles: the effect of speaking rate and phonetic context. *Journal of Speech Language, and Hearing Research*, 45, 663 – 675.

Yunusova, Y., Weismer, G., Westbury, J. R., & Lindstrom, M. J. (2008). Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech Language, and Hearing Research*, *51*, 596 – 611.

Effect of neurotype identity of conversational partners on communicative success and speaking style

James Taylor[1], Melissa A. Redford[2]

[1]University of Oregon
[2]University of Oregon
jameyt@uoregon.edu, redford@uoregon.edu

Introduction. A deficit model of Autism has limitations (e.g., Baron-Cohen et al. 1985), not least of which is in the interpretation of study results that disparage the humanity and agency of Autistic individuals (Gernsbacher & Yergeau 2019; Kapp 2019; Milton 2012), perpetuate harmful stereotypes (Bottema-Beutel et al. 2021), and undermines the interests of the primary stakeholders, namely, the Autistic community (Roche et al. 2021). Damian Milton (2012) proposed the Double Empathy Problem (DEP) framework as an alternative model for understanding Autism. According to the DEP framework, the divergence between Autistic and non-Autistic (Allistic) manner of thinking leads to different communicative strategies, which engenders communication difficulty between speakers. It follows that communication between individuals with the same neurotype identity (Autistic or Allistic) will be more successful than communication between individuals with different neurotype identities. The hypothesis has received some support in a prior study that used a diffusion chain task (much like the game Telephone) to measure communicative success among Autistic and Allistic (Crompton et al. 2020a). The researchers found that chains of all Autistic or all Allistic individuals passed information down the chain equally well, but that chains composed of both Autistic and Allistic individuals performed significantly worse. The current study sought to replicate these prior findings and extend them into the speech domain by investigating the prosody of turn-taking behavior and acoustic correlates of speaking style in conversational dyads. It is often observed that Autistic individuals have atypical prosody and do not shift speaking styles in the same manner as Allistic individuals (Bonneh et al. 2011; McCann & Peppé 2010; Zhang et al. 2022). The specific new aim of the current study was to test the DEP-motivated hypothesis that prosodic and speaking style similarities between individuals of similar neurotype identity correlate with communicative success.

Methods. A cooperative spot-the-difference task (the Diapix task; Baker & Hazan 2011) was used to elicit sustained one-on-one communication between self-identified Autistic and Allistic individuals (see McDonald 2020) under two conditions: a 'self-same' condition and a 'mixed' condition: if a participant identified as Autistic, they participated in the self-same Autistic dyad condition; it not, they participated in the self-same Allistic dyad condition; all participated in the Mixed dyad condition. Data has so far been collected from 24 speakers, resulting in 6 Autistic dyads, 6 Allistic dyads, and 12 Mixed dyads. Another 12 speakers are scheduled to participate in the study after winter break. Data collection procedures were as follows: participants completed 3 different spot-the-difference scenes per condition on different days, with the order of the conditions counterbalanced across participants. Participants were given 10 minutes per scene to find the 12 differences in each. At the end of the task, participants completed a rapport reflection (from Rifai et al. 2022). Each participant also completed a recorded reading task connected to the spot-the-difference scenes, each of which elicits a fixed set of target words. The reading task included a manipulation to elicit the target words, embedded in sentences, once in a casual speaking style and once in a clear speaking style.

Each participant produces approximately 36 minutes of conversational speech and 25 minutes of read speech. For each dyad, we have measures of transaction times, number of items found, and self-reported rapport reflections. At this point, the conversational speech has been automatically transcribed using Google Cloud Speech-to-Text, which also provided rough timestamps aligned to word boundaries. The text files have been converted to Praat (Boersma & Weenik 2023) TextGrid files and are being manually corrected. The next steps in the analyses are to extract temporal and fundamental frequency measures to investigate prosodic differences across speakers and dyads. We will also be extracting relative vowel duration and amplitude as well as spectral measures for all full vowels in the target words obtained during conversational and read speech tasks in order to investigate the effect of neurotype identity on speaking style differences.

Results. Very preliminary analyses of the data collected so far appear to contradict the basic DEP hypothesis. Although no significant differences have yet been detected, there are clear trends in the data to suggest that communication was most successful in Allistic dyads, followed by Mixed dyads, followed by Autistic dyads. These patterns are evident in **Figure 1**.

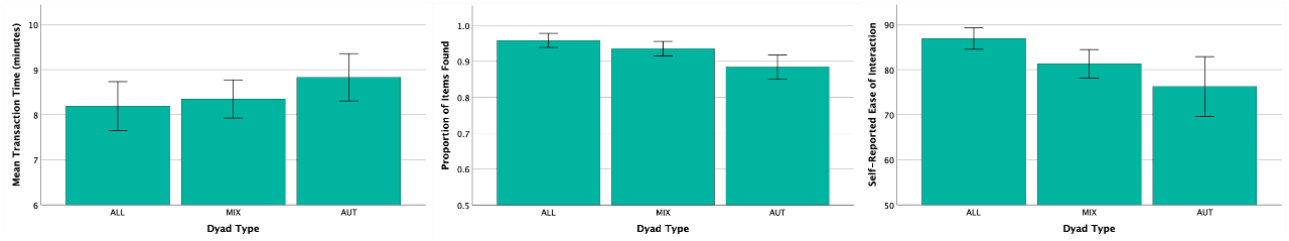


Figure 1: *Preliminary pattern of results on communicative success (left: transaction time; middle: items found; right: ease of interaction). Error bars show the 95% confidence interval.*

Discussion. Thus far we have been unable to replicate the basic finding that communication among Autistic individuals is at least as good, if not more successful, than communication among Allistic individuals. More intriguingly, the initial pattern of results suggests that communication among Autistic individuals may be less successful than between Allistic and Autistic individuals (i.e., Mixed condition): Autistic dyads took longer on average than other dyads to complete the spot-the-difference task, they found fewer of the target items in the time allotted to them, and they rated the rapport with their interlocutor lower than individuals in Mixed or Allistic dyads. With additional data, we expect to find an effect of condition on these measures. With the planned speech measures, we expect to be able to provide an initial explanation for the effect. While the overall goal of the research is to help support Autistic individuals in communicating with others, it also provides a novel, empirically-rigorous test of the validity of the DEP model in studies of speech behavior.

References

Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. Behavior Research Methods, 43(3), 761–770. https://doi.org/10.3758/s13428-011-0075-y

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition, 21(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Boersma, Paul & Weenink, David (2023). Praat: doing phonetics by computer [Computer program]. Version 6.4.01, retrieved 30 November 2023 from http://www.praat.org/

Bonneh, Y. S., Levanon, Y., Dean-Pardo, O., Lossos, L., & Adini, Y. (2011). Abnormal Speech Spectrum and Increased Pitch Variability in Young Autistic Children. Frontiers in Human Neuroscience, 4. https://doi.org/10.3389/fnhum.2010.00237

Bottema-Beutel, K., Kapp, S. K., Lester, J. N., Sasson, N. J., & Hand, B. N. (2021). Avoiding Ableist Language: Suggestions for Autism Researchers. Autism in Adulthood, 3(1), 18–29. https://doi.org/10.1089/aut.2020.0014

Crompton, C. J., Ropar, D., Evans-Williams, C. V., Flynn, E. G., & Fletcher-Watson, S. (n.d.). Autistic peer-to-peer information transfer is highly effective.

Crompton, C. J., Sharp, M., Axbey, H., Fletcher-Watson, S., Flynn, E. G., & Ropar, D. (2020). Neurotype-Matching, but Not Being Autistic, Influences Self and Observer Ratings of Interpersonal Rapport. Frontiers in Psychology, 11, 586171. https://doi.org/10.3389/fpsyg.2020.586171

Gernsbacher, M. A., & Yergeau, M. (2019). Empirical failures of the claim that autistic people lack a theory of mind. Archives of Scientific Psychology, 7(1), 102–118. https://doi.org/10.1037/arc0000067

Kapp, S. K., Steward, R., Crane, L., Elliott, D., Elphick, C., Pellicano, E., & Russell, G. (2019). 'People should be allowed to do what they like': Autistic adults' views and experiences of stimming. Autism, 23(7), 1782–1792. https://doi.org/10.1177/1362361319829628

McCann, J., & Peppé, S. (2003). Prosody in autism spectrum disorders: A critical review. International Journal of Language & Communication Disorders, 38(4), 325–350. https://doi.org/10.1080/1368282031000154204

Milton, D. E. M. (2012). On the ontological status of autism: The 'double empathy problem.' Disability & Society, 27(6), 883–887. https://doi.org/10.1080/09687599.2012.710008

Rifai, O. M., Fletcher-Watson, S., Jiménez-Sánchez, L., & Crompton, C. J. (2022). Investigating Markers of Rapport in Autistic and Nonautistic Interactions. Autism in Adulthood, 4(1), 3–11. https://doi.org/10.1089/aut.2021.0017

Roche, L., Adams, D., & Clark, M. (2021). Research priorities of the autism community: A systematic review of key stakeholder perspectives. Autism, 25(2), 336–348. https://doi.org/10.1177/1362361320967790

Zhang, M., Xu, S., Chen, Y., Lin, Y., Ding, H., & Zhang, Y. (2022). Recognition of affective prosody in autism spectrum conditions: A systematic review and meta-analysis. Autism, 26(4), 798–813. https://doi.org/10.1177/1362361321995725

Palatalisation in Russian fricatives - ISSP 2024

Natalja Ulrich[1], Jalal Al-Tamimi[2]

[1] *University of Oulu, Finland*

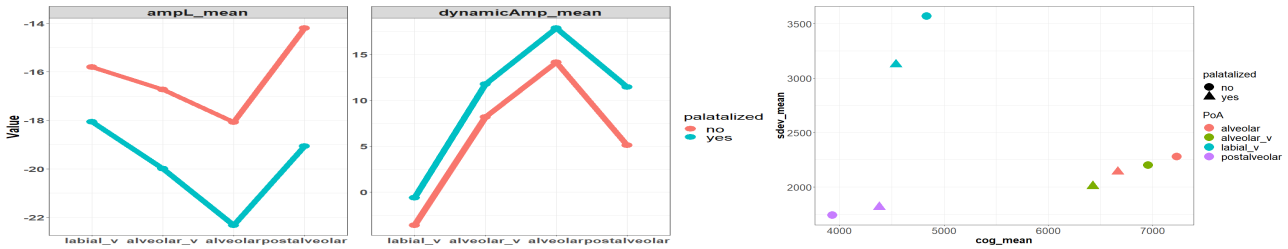
[2] *Université Paris Cité, CNRS, Laboratoire de linguistique formelle, F-75013 Paris, France*

natalja.ulrich@oulu.fi, jalal.al-tamimi@u-paris.fr

Introduction. The current study presents the results of an extended set of acoustic measures, designed to capture the underlying articulatory and aerodynamic conditions of plain and palatalized fricatives of different places of articulation. *Palatalization* refers to various articulatory processes and can denote both secondary articulation and a suprasegmental feature. The presence of a palatalization contrast in a language is contingent upon various parameters, primarily determined by the place (and secondarily, the manner) of articulation of the consonant itself (Timberlake 2004). Even if palatalization is in general cross-linguistically common, there are only a few languages contrasting more than two phonemic sibilant fricatives as the Russian language (Maddieson et al. 2013). The limited occurrence of palatalized fricatives can be attributed to the inherent synchronic and diachronic instability associated with secondary articulation contrasts, having the potential to complicate the articulatory and acoustic structures observed in fricatives. Furthermore, palatalization is similar but not identical for sounds of different places of articulation (PoA) and articulatory and acoustic properties can vary in combination with different vowels and in consonant-cluster (Timberlake 2004). In the pursuit of defining specific acoustic characteristics for plain fricatives of different PoA, research highlights the importance of the first and second spectral moments as distinctive parameters, particularly in the context of sibilant fricatives (e.g. Forrest et al. 1988; Maniwa, Jongman, and Wade 2009). Spectral moments and duration also served often to measure palatalization contrast in several languages including Romanian (Spinu, Vogel, and Timothy Bunnell 2012), Polish (Lorenc et al. 2022), Russian (Kochetov 2017; Spinu, Kochetov, and Lilley 2018), Japanese and German (Tronnier and Dantsuji 2008).
The present study adds to the exploration of distinct acoustic characteristics between plain and palatalized fricatives employing an extended set of measures. The objective is to identify unique acoustic properties and assess whether these properties can explain variations in fricative pairs with different places and manners of articulation. This inquiry aims to shed light on whether palatalization constitutes an overarching phonological feature independent of PoA.

Methods. The sample of 9070 tokes ([v]-[v^{j}], [z]-[z^{j}], [s]-[s^{j}], [ʃ]-[ɕ]), recorded from 59 native Russian speakers, is taken from a fricative dataset (Ulrich 2023b). Details of the participants, experimental design, recording process, stimuli, segmentation, and annotation can be found in Ulrich (2023a). This study looks for the first time at the palatalization contrast in Russian using an extensive set of acoustic measures of the noise portion to identify potential systematic changes that can pinpoint to an overarching phonological patterning in this large-scale dataset.
For acoustic analyses, each fricative was divided into 11 time frames to track the dynamic change associated with palatalization. Then, 10ms intervals within each time frame were extracted, from which 256-point DFT spectra without zero-padding, with a bin width of 80Hz, were obtained. Acoustic measures similar to those of Al-Tamimi and Khattab (2015) consisted of an extended set in the frequency range up to 20500 Hz, including the standard parameters: peak frequency (*peakHz*) and amplitude (*peakAmpdB*), spectral moments (centre of gravity (*cog*), standard deviation (*sdev*), kurtosis (*kurt*), skewness (*skew*)), duration (*dur*), and number of zero crossing (*zrp*), and the amplitude measures: amplitude low (*ampL*), high (*ampH*), dynamic amplitude (*dynamicAmp*), mean amplitude in low (*ampLMin*) and mid (*ampMid*) frequency ranges and sibilance (*sibilance*). We used Random Forest, a Machine Learning classifier, and mean values to evaluate distinct parameters of plain and palatalized fricatives. Our study comprised three prediction tests: individual fricatives, palatalization in general, and binary classification by PoA.

Results. The classification accuracy is similar for the three different methods. Plain sibilants [s] and [ʃ] show an accuracy of around 98%, while only half of [s^{j}], and [ʃ^{j}] are correctly classified, the other half is predicted as plain fricatives. For the plain voiced bilabial and alveolar fricatives, we observe similar patterns. The classification rate for [v^{j}], is around 60% and



(a) Low amplitude (*ampL*) and dynamic amplitude (*dynamicAmp*)

(b) *cog* by *sdev* space

36% are miss-classified, while most [z^j] tokens are classified as plain fricatives. The variable importance differs slightly between the three sets. In predicting the single fricatives most important variables are *zcp*, *cog*, *dynamicAmp*, and *peakHz*. For the classification of plain vs. palatalized fricative, the most crucial measures are *dynamicAmp*, *ampL*, *cog*, and *peakHz*. The pairwise comparison by PoA suggests *ampL*, *dynamicAmp*, and *ampLMin* as the best parameters. Figure 1a shows the patterns across the best two amplitude-based measures: *ampL* (frequencies 0-2000Hz) and *dynamicAmp*. The former displays an overall decrease in the amplitude at the low frequencies, likely to be correlated with the preparation of an increase in vocal effort, highlighted by an increase in *dynamicAmp*. The patterns observed in Figure 1b indicate a clear impact of palatalization on how each place is realized: front places (e.g., bilabial, alveolar voiced and voiceless) show a reduced *cog* and *sdev*, which are indicative of a retracted PoA (↓*cog*) to a palatal place and with a more apical type of production (↓*sdev*). An opposite pattern is observed for the post-alveolar place, with increased *cog* and reduced *sdev*. The results correlated with a palatal place of articulation (↑*cog*) and a more pronounced apical type of production (↓*sdev*). These patterns are in line with previous findings (e.g. Maniwa, Jongman, and Wade 2009).

Discussion. The results indicate that palatalization serves as a phonological feature in Russian fricatives. However, due to significant misclassification rates, with almost half of palatalized fricatives being misclassified as plain, it cannot be concluded that palatalization is uniformly implemented across all fricatives. Amplitude-based measures, particularly in binary comparisons by PoA, outperform other measures in discriminating palatalization contrast. These findings imply heightened excitation strength and increased effort during production represented mainly by the dynamic amplitude (*dynamicAmp*), in addition to a change in the constriction location and the portion of the tongue used during their production. Further interpretation reveals that palatalization is not consistently implemented, and the observed variation may stem from linguistic context effects or inter-speaker differences, which needs further investigation.

References.

Al-Tamimi, Jalal and Ghada Khattab (July 2015). "Acoustic cue weighting in the singleton vs geminate contrast in Lebanese Arabic: The case of fricative consonants". en. In: *The Journal of the Acoustical Society of America* 138.1, pp. 344–360. DOI: 10.1121/1.4922514.

Forrest, Karen, Gary Weismer, Paul Milenkovic, and Ronald N. Dougall (July 1988). "Statistical analysis of word-initial voiceless obstruents: Preliminary data". en. In: *The Journal of the Acoustical Society of America* 84.1. Number: 1, pp. 115–123. DOI: 10.1121/1.396977.

Kochetov, Alexei (2017). "Acoustics of Russian voiceless sibilant fricatives". In: *Journal of the International Phonetic Association* 47.3, pp. 321–348. DOI: https://doi.org/10.1017/S0025100317000019.

Lorenc, Anita, Marzena Zygis, Daniel Pape, Lukasz Mik, and Marton Soskuthy (2022). "Articulatory and acoustic variation in Polish palatalised retroflexes compared with plain ones". In: *Journal of Phonetics*. DOI: 10.1016/j.wocn.2022.101181.

Maddieson, Ian, Sébastien Flavier, Egidio Marsico, Christophe Coupé, and François Pellegrino (Aug. 2013). "LAPSyd: lyon-albuquerque phonological systems database". en. In: *Interspeech 2013*. ISCA, pp. 3022–3026. DOI: 10.21437/Interspeech.2013-660.

Maniwa, Kazumi, Allard Jongman, and Travis Wade (June 2009). "Acoustic characteristics of clearly spoken English fricatives". en. In: *The Journal of the Acoustical Society of America* 125.6. Number: 6, pp. 3962–3973. DOI: 10.1121/1.2990715.

Spinu, Laura, Alexei Kochetov, and Jason Lilley (June 2018). "Acoustic classification of Russian plain and palatalized sibilant fricatives: Spectral vs. cepstral measures". en. In: *Speech Communication* 100, pp. 41–45. DOI: 10.1016/j.specom.2018.04.010.

Spinu, Laura, Irene Vogel, and H. Timothy Bunnell (Jan. 2012). "Palatalization in Romanian—Acoustic properties and perception". en. In: *Journal of Phonetics* 40.1, pp. 54–66. DOI: 10.1016/j.wocn.2011.08.001.

Timberlake, Alan (2004). *A Reference Grammar of Russian*. en. Cambridge University Press.

Tronnier, Mechtild and Masatake Dantsuji (2008). "Some Acoustic Characteristics of Glottal and Palatal Fricatives in Japanese and German". en. In: *Selected Proceedings of the 4th Workshop on Spanish Sociolinguistics*, pp. 54–63.

Ulrich, Natalja (June 2023a). "Database description: Russian fricatives recorded in 198 real speech sentences from 59 speakers". en. In: *Data in Brief* 48, p. 109205. DOI: 10.1016/j.dib.2023.109205.

— (2023b). *Russian Fricatives [Dataset]*. en. Version Number: 1.0 Type: dataset. DOI: 10.48656/4Q9C-GZ16.

Perception of Accentual Phrase Boundaries in L1 and L2 French

Caroline L. Smith[1], Bruno Pinto Silva[1]

[1]Department of Linguistics, University of New Mexico
caroline@unm.edu, bpslinguist@unm.edu

Introduction. This study examines prosodic phrasing in French noun phrases, specifically the differences in how L1 and L2 speakers realize it. The basic unit of phrasing in French is the Accentual Phrase (AP) (Jun & Fougeron 2002). Although in most cases each content word is in a separate AP, noun phrases with short nouns and adjectives are optionally produced with both content words in the same AP (Post 1990). Potential variation in the parsing of APs is tested here by varying the length of the noun and the presence/absence and length of the post-nominal adjective. Acoustically, Accentual Phrases in French are characterized by lengthening on the final syllable (Pasdeloup 1990), distinguishing AP-final and AP-medial words. To our knowledge the most extensive investigation of AP length is Delais-Roussarie (1996), who found that seven syllables was the most in a single AP. No previous work has systematically investigated AP parsing in L2 speakers. Here we use listener perceptions as well as acoustic measures to "measure" AP length.
Our hypotheses are as follows:

H1: When the total number of syllables in the AP exceeds seven, the noun and adjective will be perceived as belonging to separate APs: listeners will consistently perceive a boundary between the noun and adjective.
H2: L1 listeners will perceive more boundaries in L2 speech because the L2 speakers produce acoustic cues to AP-final position on all content words, over-generalizing the pattern whereby each content word is in a separate AP.

Methods. The study consisted of two phases: recording L1 and L2 speakers of French producing utterances that included noun phrases of varied lengths, then testing L1 French listeners' perceptions of the recorded speech.
All speakers were recorded in Paris. The twelve L1 French speakers (all female) were recruited in undergraduate classes at a Paris university. The twelve L2 speakers (three male) were recruited through email lists and flyers. All had American English as their L1, and were initially exposed to French through classroom instruction in the US between the ages of 12 and 19, but were resident in Paris at the time of recording. The criterion for inclusion was that their French should be at least C1 (advanced) level in the Common European Framework.
All speakers read the same material, which consisted of three-sentence paragraphs. The target noun phrase occurred phrase-medially in the second sentence. Sets of five paragraphs were constructed such that the second sentence was identical in all five except that in one sentence there was no adjective, and in the other four sentences in the set, there were post-nominal adjectives canonically pronounced with one, two, three or four syllables. The noun phrases were either the direct object of the sentence or the object of a prepositional phrase in the predicate. Here is a sample sentence with the target noun phrase underlined: *Elle boit toute une bouteille <u>d'eau minérale</u> d'un seul coup.* 'She drinks a whole bottle of mineral water all at once.' (This illustrates a monosyllabic noun, *eau*, with a three-syllable adjective, *minérale*.)
In order to facilitate f0 tracking, the nouns were all entirely voiced, as were the adjectives with three exceptions. Also, except for *d'eau*, each noun was preceded by a monosyllabic function word within the same noun phrase, which resulted in phrases ranging from one to nine syllables. Another monosyllabic noun *vin*, 'wine', was also used, plus nouns of two (*melon*, 'melon'), three (*débardeur*, 'tank top'), and four (*désagrément*, 'inconvenience') syllables. This resulted in five sets of paragraphs and a total of 25 paragraphs. In the materials read by the speakers, these were interspersed with 20 unrelated paragraphs. Three different random orderings were created, and all speakers read them in the same order.
The recorded paragraphs were segmented manually in Praat (Boersma & Weenink 2022), then a script automatically extracted the durations of the nouns and adjectives, plus the median f0 of each syllable.
The perception portion of the study used Rapid Prosody Transcription (RPT), a method of obtaining listener interpretations of prosody in (close to) real time (Cole et al. 2017). This was accomplished using purpose-designed software LMEDS (Mahrt 2016), a web-based platform that enables RPT experiments to be run over the Internet. Participants hear recorded speech while reading on-screen the text corresponding to that speech (which is transcribed without punctuation, and with capitalization restricted to proper names). They are asked to click on the screen to indicate phrasal boundaries: the instructions are to listen for words followed by "a rupture or discontinuity in the flow of speech" ("une rupture ou une discontinuité dans le flux de la parole"), and to click between the words where they perceive this break. A vertical bar | appears on the screen between the words. Participants have the option of listening to the speech a second time and can change their markings. Recordings of L1 and L2 speakers were presented in distinct blocks.
59 L1 users of French were recruited from undergraduate classes at a university in Lyon. They participated remotely, via a link to the web site running LMEDS. For each of the 25 distinct paragraphs, listeners responded to one recording from an L1 speaker and one from an L2 speaker, in separate blocks.

The listener responses for each word were converted to 'b-scores', equal to the proportion of listeners who marked the word as being followed by a boundary. The b-scores for the nouns and adjectives were the dependent variables in two generalized linear models (glm in R) that were fit to the data. The fixed factors were **speakerGroup** (L1 or L2), **numNounSylls** (1 - 4), the number of syllables in the noun, and **numAdjSylls** (0 – 4), the number of syllables in the adjective, if any. The model also included all interactions among these factors, but no random factors since each speaker was heard only two or three times in the set of recorded paragraphs used for the listening task.

Results. Listener agreement was assessed using Fleiss's kappa over the total of 1612 words in the recordings included in the perception study. For the responses to L1 speakers, kappa was 0.69, slightly higher than the 0.62 agreement for the listener responses to L2 speakers. These values are in the range described as "substantial agreement" by Landis & Koch (1977), and are similar to values obtained in other RPT studies (e.g., Cole et al. 2017).
In the linear model testing structural factors, b-scores for the nouns were significantly higher for L2 speakers than L1 ($p<.01$), although numerous two- and three-way interactions showed that this was not true in all contexts (see Figure 1). Most striking is the case of 4-syllable nouns combined with 4-syllable adjectives, where the b-score for L1 speakers was 0.56 while for L2 speakers it was 0.12. Listeners clearly thought that the L1 speakers were producing a boundary after the noun in this, which was the longest noun phrase in the data set, but did not perceive a boundary in the L2 speech.
Statistical models were also run with the same main effects as before, and adding the **duration of the final syllable of the noun** and the **median f0 of the final syllable** as predictors for the model for b-score of the noun. The **duration of the final syllable** was significant ($p<.02$) in predicting the b-score, as was its **interaction with speakerGroup**: the effect of duration was limited to the L1 speakers (regression slope 1.70); it did not predict b-scores for L2 speakers (slope 0.05).

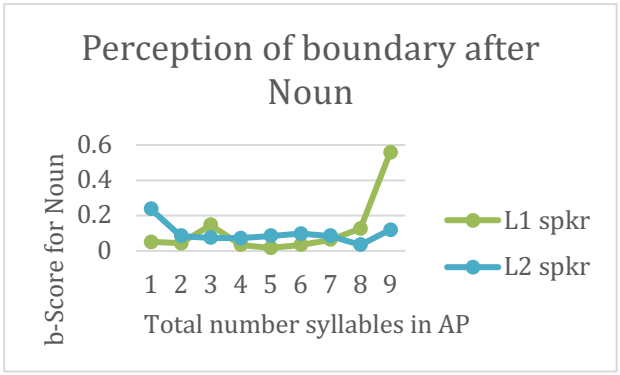


Figure 1: *Proportion of French listeners who perceived a boundary after the noun in Det Adj Noun phrases, comparing perceptions when listening to L1 or L2 speakers.*

Discussion.

H1: Not supported in its original form. For L1 speakers, L1 listeners reliably perceived a boundary between noun and adjective only when the total number of syllables in the noun phrase reached nine. This is longer than production evidence suggested in previous studies also based on reading aloud. For L2 speech, there was no maximum length at which a boundary was frequently perceived.

H2: Partially supported. Listeners did perceive more boundaries overall in L2 speech. Acoustic measurements show that the L2 speakers consistently lengthened the final syllable of the noun, which suggests that it is in a separate AP, but the durational variation did not predict the perception of additional boundaries.

Operationalizing listener perceptions with RPT can distinguish which L1-L2 differences contribute to perception differences, suggesting strategies for L2 learners to be more comprehensible to L1 listeners.

References

Boersma, P. & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program].

Cole, J., Mahrt, T., & Roy, J. (2017). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes,* 25, 1141-1177.

Delais-Roussarie, E. (1996). Phonological phrasing and accentuation in French. In M. Nespor & N. Smith (Eds.), *Dam Phonology : HIL Phonology Papers II.* Holland Academic Graphics. 1–38.

Jun, S-A., & Fougeron, C. (2002). Realizations of accentual phrase in French intonation. *Probus,* 2, 147-172.

Landis, J. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* 33, 159-174.

Mahrt, T. (2016). LMEDS: Language markup and experimental design software. https://github.com/timmahrt/LMEDS

Pasdeloup, V. (1990). *Modèle de règles rythmiques du français appliqué à la synthèse de la parole.* Ph.D. thesis, Université de Provence.

Post, B. (1990). Restructured phonological phrases in French: evidence from clash resolution. *Linguistics,* 37, 41-63.

Sexual dimorphism of vocal tract development from gestation to adulthood: Mixed-effects modelling of an extended X-ray database

Guillaume Barbier, *GIPSA-Lab, Univ. Grenoble Alpes, France*
Louis-Jean Boë, *GIPSA-Lab, Univ. Grenoble Alpes, France*
Guillaume Captier, *Anatomy Laboratory, Univ. Montpellier, France*
Rafael Laboissière, *LPNC, CNRS, Univ. Grenoble Alpes, France*

Human vocal tract growth during ontogenesis is known to be non-uniform, in the sense that various structures undergo their own growth rhythm. To understand how morphological constraints act on the development of speech production, detailed anatomical knowledge of the developing vocal tract is necessary. To this aim, longitudinal X-ray data (966 sagittal cephalometric radiographs) of 68 American people were used to quantify the anatomical development of the human supra-laryngeal vocal tract. The anatomical sections of 12 fetuses were added to ensure the continuity of data around birth. Five main anatomical structures were investigated: the hard and soft palates, the oral and pharyngeal cavities together with the estimated vocal tract length. Growth curves and rates were estimated for each variable of interest, using a double sigmoid model, which capture the essential aspects of the anatomical development of the underlying structures. The significance of the paramenters of the model were statistically evaluated using non-linear mixed-effect models, which are able to take into account the longitudinal data for each participant. Results indicate that each structure follows its own growth rhythm, but all structures present two growth spurts, one during gestation and a second during puberty, and that sexual dimorphism appears during puberty and applies mainly to vertical structures. These data are useful to understand how anatomical constraints act on speech development, to model vocal tract growth, and to better understand the articulatory-acoustic relationships during ontogenesis.

Combining manual control of intonation with whisper articulation in voice substitution: the case of contrastive focus

Delphine Charuau[1], Nathalie Henrich Bernardoni[1], Silvain Gerber[1], Olivier Perrotin[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab Grenoble, F-38000 France

delphinecharuau1@gmail.com, olivier.perrotin@grenoble-inp.fr

Introduction. The substitution of the glottal source with a synthetic one, for instance following a laryngectomy, requires a substitute control of intonation, such as pressing a button on an electrolarynx. This paradigm combines a manual control of intonation with natural articulation, which has poorly been evaluated in terms of efficiency in fulfilling linguistic functions. This speech task is made challenging by the diverse prosodic functions of the fundamental frequency (f_o) such as focus, boundary marking, attitudes, and emotions (Mertens 2008). In this study, we are particularly interested in the realisation of contrastive focus, which is characterised by an increase in the f_o curve (Jun & Fougeron 2000; Grice *et al.* 2017). Our aim is to analyse to which extent participant are able to produce the emphasising of syllables through variations in intonation controlled by manual gestures. Following the numerous developments in human-machine interfaces for speech control (d'Alessandro 2022), two complementary gestures will be compared.

Methods. Our experiment was conducted using a whisper-to-speech conversion (WSC) software (Ardaillon *et al.* 2022; Perrotin & McLoughlin 2020), which enables the real-time conversion of whispers acquired with a microphone into vocalised speech while providing control of intonation with hand gesture through human-machine interfaces. In this study, we compared two complementary control modalities: the first is isometric and allows modulation of intonation through finger pressure on a button as in the Trutone electrolarynx, while the second is isotonic and allows control through wrist rotation, similarly as beat gestures (Leonard & Cummins 2011). Both the degree of button depression and the angle of wrist rotation are linearly mapped to f_o, in semitones, in the range of an octave around the speaker's mean f_o value, measured in a calibration step.

To encourage speakers to produce a contrastive focus at a specific location but without giving any explicit instruction (Dohen 2005), they were recorded in simulated dyadic interactions guided by a scenario displayed on a screen. Speakers started with the production of an initial utterance (condition "*pre*"), followed by a pre-recorded question simulating the misunderstanding of a target word. The speaker had then to repeat the same utterance, potentially introducing a focus on the target word that was misunderstood (condition "*post*"). These interactions were based on a corpus of 6 sentences, each composed of 9 monosyllabic words (CV-type), evenly distributed among the subject, verb, and object constituents. The subject and object constituents of each sentence each contain the syllable [lu], of which only one is targeted at a time with the question that follows. The position of the [lu] syllable within the constituents (S1, S2, S3, O1, O2, O3) varied from one sentence to another, so that overall it is seen as targeted and non-targeted for all syllable positions. The interaction task was carried out in three production modes: with *Natural voice*, with *Finger pressure* control and with *Wrist movement* control. The two latter conditions use the WSC system and their order was randomly chosen across participants. Each production mode was preceded by a training phase to become familiar with the interaction scenario, and the interfaces.

Sixteen speakers were recorded (median age = 24.5 years old; Q1 = 22.5; Q3 = 27). They did not report any speech, hearing, arm, or hand motor disorders. The acoustic data was semi-automatically segmented and annotated using Astali (Loria 2016) and Praat (Boersma & Weenink 2021). Matlab was used to extract temporal (relative duration of syllables, utterance duration, articulation rate) and intonation data (height, position, and width of the centred f_o peak). Centred f_o (f_{oc}), expressed in semi-tones (st), corresponds to the subtraction of median f_o values computed for one speaker and one interface to the corresponding raw f_o. The relative duration (Dr) of syllables is expressed as a percentage of the sentence duration. Statistical analyses were conducted using R. The significance of the results was tested through a mixed-effects linear regression model to examine the effects of syllable position, interfaces, and syllable condition. Random factors such as *speaker* and *repetition* were also taken into account. The overall significance level was set to $p < 0.05$.

Results. Fig. 1 displays peak f_{oc} height per production mode and syllable position. In *Natural voice*, speakers tend to mark the focus on the target syllable by raising the f_o curve in the "*post*" condition (dark green) compared to the "*pre*" condition (dark red), regardless of the target syllable position in the sentence. However, this difference is only significant, when the target syllable is in the second position within the object constituent (O2). In contrast, in the *Finger pressure* task, this difference is significant, except when the target syllable is in the first position of a constituent. In the case of the *Wrist movement* task, the difference between the "*pre*" and "*post*" conditions of the target syllable is significant, regardless of the syllable position.

We also observed that the [lu] syllables are significantly longer when we expect a focus, both in *Natural voice* (Dr_{mean} = 16.1 ± 2.9%), *Finger pressure* (Dr_{mean} = 18.1 ± 4.5%) and *Wrist movement* (Dr_{mean} = 18.3 ± 4.4%), than when they do not

(*Natural voice:* Dr_{mean} = 12.3 ± 1.9%; *Finger pressure*: Dr_{mean} = 12.1 ± 3.1%; *Wrist movement:* Dr_{mean} = 12.3 ± 2.9%), regardless of the syllable position in the utterance.
Finally, we analysed the position of the f_o peak relatively to the target syllable boundaries (Pos) in the *"post*" condition. In *Natural voice*, the f_o peak tended to be located towards the end of the marked syllable (Pos_{mean} = 70.3 ± 27.5%). When using an interface for f_o control, the f_o peak was achieved slightly earlier during the production of the target syllable, i.e., the peak being more centred relatively to the boundaries of the syllable (*Finger pressure:* Pos_{mean} = 42.7 ± 35%; *Wrist movement:* Pos_{mean} = 51.8 ± 36.1%).

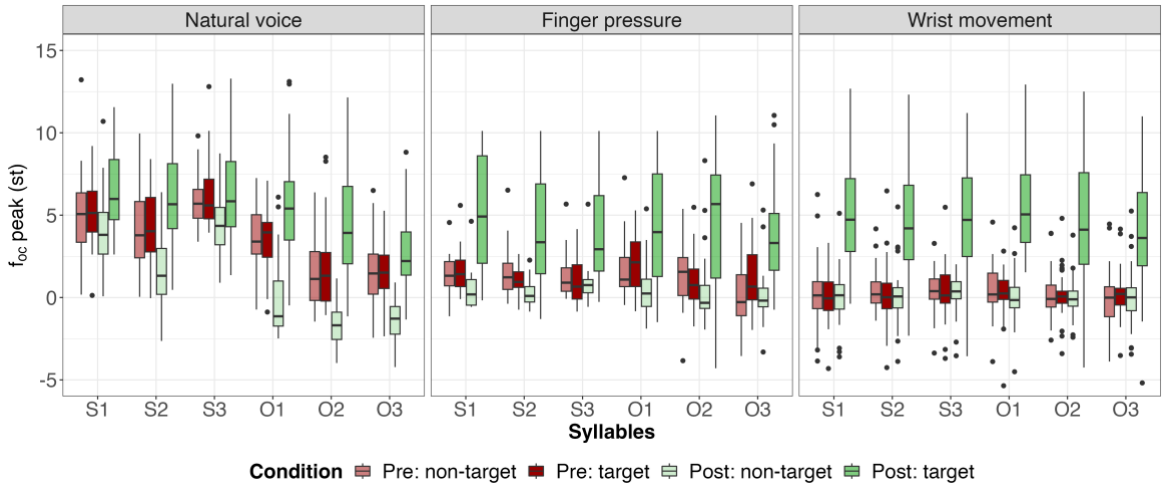


Figure 1: *f_{oc} peak height of the [lu] syllables according to their location on the utterance.*

Discussion. Explicit manual control of intonation requires to become aware of one's own intonation curve in speech, which is usually implicit in typical speech production. The question of the difficulty of external manual f_o control to realise a specific linguistic function was therefore not trivial. However, all speakers were able to successfully produce an elicited contrastive focus in a paradigm of external and explicit intonation control, within the relatively limited time of this experiment (one hour).
In *Finger pressure* and *Wrist movement* tasks, increased f_o on [lu] focused syllables demonstrated: i) the speakers' intention to distinguish this syllable from the others by modulating intonation, and ii) their awareness of the important role of its function for emphasising the target syllable. Focus realisation was also marked by a significant lengthening of the relative duration of the target syllable, both in *Natural voice* and in whispered speech with manual intonation control, regardless of articulation rate. Control of intonation with hand gesture was synchronised with syllable production: if the f_o peak was reached earlier than *Natural voice*, it was mostly realised within the boundaries of the [lu] target syllable. More specifically, in *Wrist movement* task, the f_o peak gravitated from the centre of syllable, regardless of the syllable position, while it was slightly anterior in *Finger pressure* task. The comparison of interface usage showed no significant differences between these two types of control, although we observed some specificities in their use, which will be investigated in future work. These encouraging results call for the exploration of other linguistic functions in a less controlled speech task, to fully validate such control paradigm in voice substitution applications.

References

Ardaillon, L., Henrich Bernardoni, N., & Perrotin, O. (2022). Voicing decision based on phonemes classification and spectral moments for whisper-to-speech conversion. In *Interspeech 2022*, 2253-2257.

Boersma, P., & Weenink, D. (2021). *Praat: Doing by computer* [Computer program], Version 6.1.42.

d'Alessandro C. (2022). Une nouvelle organologie de la voix : chironomie et prosodie de la parole et du chant. *Actes des Journées d'Etudes sur la Parole,* 625-636.

Dohen, M. (2005). Deixis prosodique multisensorielle : production et perception audiovisuelle de la focalisation contrastive en français. *Thèse de doctorat, Institut National Polytechnique de Grenoble*.

Grice, M., Ritter, S., Niemann, H., & Roettger, T. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics, 64*, 90-107.

Jun, S.-A., & Fougeron, C. (2000). A phonological model of French intonation. In Botinis, A. (Ed.), Intonation: Analysis, Modeling and Technology, Dordrecht: Kluwer Academic. 209-242.

Laboratoire lorrain de recherche en informatique et ses applications – UMR 7503 (Loria). (2016). ASTALI [Outil]. *ORTOLANG (Open Resources and TOols for LANGuage)* – www.ortolang.fr, v2.

Leonard T. & Cummins F. (2011). The temporal relation between beat gestures and speech. *In Language and Cognitive Processes*, 26(10), 1457-1471.

Martens, P. (2008). Syntaxe, prosodie et structure informationnelle : une approche prédictive pour l'analyse de l'intonation dans le discours. *Travaux de linguistique,* 56, 97-124.

Perrotin, O., & McLoughin, I. (2020). Glottal flow synthesis for whisper-to-speech conversion. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 889-900.

Laryngeal movements in the production of French stops, with variations in phonation mode and intensity

Maëva Garnier, Myriam Fantone, Nathalie Henrich Bernardoni

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
maeva.garnier@gipsa-lab.fr

Introduction. Functional dysphonia is characterized by a hyperfunctional phonatory behavior, leading to increased laryngeal fatigue, and possible lesions on the vocal folds (Hillman et al. 1990). So far, this vocal hyperfunction has mainly been characterized on sustained vocalizations, from the conversion ratio of subglottal pressure and oral airflow into acoustic energy ("vocal efficiency"), the degree of vocal fold adduction, or the vocal fold vibration pattern (Gauffin and Sundberg, 1989; Aronsson et al. 2007). A range of inefficient or potentially harmful vocal behaviors has also been identified and grouped under the terminology "Muscle Misuse Voice Disorders" (Morrison and Rammage, 1993), extending the characterization of vocal hyperfunction to more dynamic aspects of glottal behavior (voicing onsets (Andrade et al. 2000)), to the articulatory dimension of laryngeal behavior (vertical movements (Iwarsson and Sundberg, 1998)) and even to non-laryngeal aspects of phonation behavior (posture, breathing patterns, ... (Giovanni et al. 2008)). As pointed out by Sama et al. (2001), laryngeal dynamic of dysphonic patients is much better assessed during connected speech. An important source of laryngeal neuromuscular fatigue and vocal fold microtrauma may not only come from vocal fold vibration, but also in large part from glottal onsets after each voicing interruption, and from dynamic movements of the larynx, as involved in the production of voiced stop consonants.
The first objective of the present study is to characterize laryngeal movements involved in the production of stop consonants in adults with no speech or voice disorders, in order to reference their "typical" coordination and their variation as a function of vocal effort. The medium-term objective is to compare these typical movements with those of dysphonic patients, in order to understand the influence of laryngeal dynamics to vocal hyperfunction.

Methods. The study relies on endoscopic videos of the larynx that were recorded from a French female speaker, while she produced 6 repetition of the syllables [Ca] with C={p, t, k, b, d, g} in modal voice (with 3 levels of vocal effort, from murmur to shout) and whispered voice (with 2 levels of vocal effort : soft and loud whisper). The endoscopic images were recorded by an ENT doctor using flexible laryngoscopy (30 frames/s) synchronously with audio signal.
For each production, the burst interval and instants of voicing onset and offset were manually annotated in Praat from the audio signal, while the endoscopic videos were annotated in ELAN, identifying the onset and termination of different laryngeal movements: glottal adduction or abduction; ventricular narrowing or widening; epiglottis posterior tilting. Several distances between anatomical structures were measured at onset and at termination of laryngeal movements: between corniculate cartilages, between ventricular folds, and between epiglottis edge and pharyngeal cavity. The angle between the two free edges of the vocal folds was also measured.

Results. Different movements of vocal fold adduction/abduction were observed, with variable timing, depending on consonant voicing ([p] vs. [b]; [t] vs. [d]; [k] vs. [g]) and phonation mode (modal or whispered) (see Figure 1, left panel). For all consonants, a vocal fold adduction was evidenced before the occlusion onset. For consonants with an unvoiced occlusion phase (voiceless stops and whispered consonants), this initial adduction was followed by a vocal fold abduction starting in the middle of the occlusion phase, possibly linked to the build-up of intra-oral pressure. For syllables in modal voice with initial voiceless stops, the voicing onset after occlusion release was accompanied by a rapid movement of vocal fold adduction (see Figure 1, left panel). This movement was not progressive or did not anticipate the voicing onset, as for voiced stops. Even for whispered consonants (no voicing before or after occlusion), a movement of vocal fold adduction was observed after the occlusion release. Yet, it did not coincide with vowel onset as could be expected, but it started much later, almost at the middle of the vowel with no link to any visible feature on the spectrogram.
Finally, a movement of laryngeal abduction was observed for all syllables before the end of the vowel.
In all cases, the occlusion phase of stop consonants was always produced with a certain degree of arytenoid adduction, which was greater for voiced consonants (even whispered ones), increased with vocal effort in modal voice and decreased in whispered voice. Furthermore, the adduction movement that accompanies voicing onset after the occlusion release of voiceless stops became more rapid with increasing vocal effort (meaning a "harder" glottal attack) and ampler as the consonant place of articulation moved backwards (/p/</t/</k/) (see Figure 1, right panel (a)).

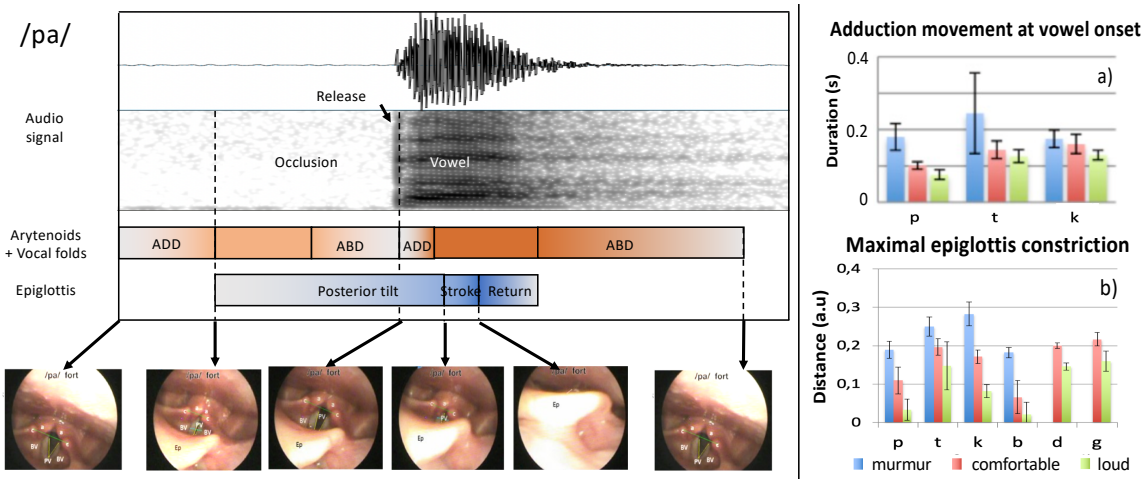


Figure 1. *Left panel: Laryngeal movements observed during the different phases of a syllable /pa/ produced in modal voice and loud intensity (ADD: adduction; ABD: abduction). Right panel: (a) duration of the movement of laryngeal adduction movement accompanying voicing onset, for voiceless stops and (b) minimum distance between the edge of the epiglottis and the pharyngeal cavity after the occlusion release.*

In parallel with these adduction/abduction movements, three-phase movement of the epiglottis was found, with 1) a progressive posterior tilt from the beginning of the syllable, followed by 2) a very rapid backward tilt (Stroke) around 50ms after occlusion release, and 3) a progressive forward return. The amplitude of epiglottis backward tilting increased with vocal effort and varied significantly with place of articulation (/p/ < /k/~/t/) (see Figure 1; right panel (b)).
No movement of arytenoids, epiglottis or ventricular folds was observed at the very moment of occlusion release.

Discussion. We can question the neuromuscular effort of these laryngeal movements, and the microtrauma they can cause on the vocal folds. During CV syllable production, voiceless stops induce several movements of laryngeal adduction/abduction/adduction, whereas voiced stops remain in adduction state. Producing voiceless stops may thus induce additional neuromuscular fatigue on the long term. Our results also showed that the occlusion phase of stop consonants is always produced with a degree of adduction, even for voiceless stops or whispered consonants. This adduction increases with vocal effort in modal voice, possibly causing greater neuromuscular fatigue. For whispered voiced stops, vocal fold adduction is accompanied by an additional adduction movement of the ventricular folds, whose narrowing increases with vocal effort. This shows how whispered phonation is not a "restful" phonation mode. During an episode of vocal fatigue, or after laryngeal surgery, murmured (modal) phonation should therefore be preferred to whispered voice. Another interesting observation is the rapid movement of vocal fold adduction observed at voicing onset for voiceless stops, whose amplitude and speed increase with vocal intensity. Abrupt and hard glottal attacks are known to cause microtrauma on the vocal folds, which can then lead to lesions on the long term (Andrade et al. 2000). This adduction movements at voicing onset of voiceless stops when producing CV syllables will therefore be one of the first aspects to be examined when we will characterize, in a second step, the hyperfunctional behavior of dysphonic patients. Finally, our data did not include any measurement of laryngeal vertical movements, which may also occur during the production of stop consonants and may also be an important source of neuromuscular fatigue. Future work should explore this movement, particularly during the occlusion phase and at the moment of occlusion release.

References

Andrade, D. F., Heuer, R., Hockstein, N. E., Castro, E., Spiegel, J. R., and Sataloff, R. T. (2000). "The frequency of hard glottal attacks in patients with muscle tension dysphonia, unilateral benign masses and bilateral benign masses," J. Voice, 14, 240–246.

Aronsson, C., Bohman, M., Ternström, S., and Södersten, M. (2007). "Loud voice during environmental noise exposure in patients with vocal nodules," Logoped. Phoniatr. Vocol., 32, 60–70

Gauffin, J., & Sundberg, J. (1989). Spectral correlates of glottal voice source waveform characteristics. Journal of Speech, Language, and Hearing Research, 32(3), 556-565.

Giovanni, A., Akl, L., and Ouaknine, M. (2008). "Postural dynamics and vocal effort: preliminary experimental analysis," Folia Phoniatr. Logop., 60, 80–85.

Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., & Vaughan, C. (1990). Phonatory function associated with hyperfunctionally related vocal fold lesions. Journal of Voice, 4(1), 52-63.

Iwarsson, J., and Sundberg, J. (1998). "Effects of lung volume on vertical larynx position during phonation," J. Voice, 12, 159–165.

Morrison, M. D., and Rammage, L. A. (1993). "Muscle misuse voice disorders: description and classification," Acta Otolaryngol., 113, 428–434.

Sama, A., Carding, P. N., Price, S., Kelly, P., & Wilson, J. A. (2001). The Clinical Features of Functional Dysphonia. The Laryngoscope, 111(3), 458-463.

Day 3

Thursday, May 16

08:00am		03:00pm	Poster Session 3
08:30am	**María Florencia Assaneo**	03:30pm	
09:00am		04:00pm	
09:30am	Oral Session 7 **Coordination I**	04:30pm	
10:00am		05:00pm	Coffee Break
10:30am	Coffee Break	05:30pm	Oral Session 10 **Methodology**
11:00am	Oral Session 8 **Phonetics/Phonology II**	06:00pm	
11:30am		06:30pm	
12:00am	Oral Session 9 **Development**	07:00pm	
12:30am		07:30pm	Dinner
01:00pm	Lunch Break	08:00pm	
01:30pm			
02:00pm			
02:30pm			

Oral session 7

Coordination I

9:30– 10:30 am

	Title	Authors
9:30 - 9:50 am	Temporal coordination of articulatory and respiratory events during utterance - initial and inter-speech pauses	Oksana Rasskazova (Humboldt- Universität zu Berlin)*; Susanne Fuchs (zas); Christine Mooshammer (Humboldt-Universität zu Berlin)
9:50 - 10:10 am	Dialect specific patterns of gestural timing? Evidence from lateral clusters.	Emily Gorman (Lancaster University)*
10:10 - 10:30 am	Phase-locking of articulation and spectral flow	Jessica A Campbell (University of Southern California)*; Louis Goldstein (University of Southern California); Leonardo Lancia (Laboratorie Parole et Langage, Aix-Marseille Université; Laboratoire de Phonétique et Phonologie, Université Sorbonne Nouvelle)

Temporal coordination of articulatory and respiratory events during utterance - initial and inter-speech pauses

Oksana Rasskazova[1], Christine Mooshammer[1], Susanne Fuchs[2]

[1]Humboldt-Universität zu Berlin, Germany
[2]Leibniz-Centre General Linguistics, Germany
oxanarass@gmail.com, mooshamc@hu-berlin.de, fuchs@leibniz-zas.de

Introduction. The results of previous studies on temporal aspects of speech planning show that the movements of the articulators usually start before the acoustic onset. Depending on the manner of articulation of the initial segment, this delay varies between 120 -180 ms (Mooshammer et al. 2012). Respiration may play a crucial role in the temporal organization of the speech preparatory activities. Indications for this relationship are coming from studies on respiratory activities prior to the utterance (Fuchs et al. 2013). However, the coordination between breathing and articulatory preparation has rarely been discussed yet. In Rasskazova et al. (2019) we presented the first results on the coordination of respiratory, acoustic, and articulatory events prior to the utterance for the upcoming alveolar consonants /t/ and /n/ for six speakers. We found evidence for temporal alignment between oral articulators and the onset of exhalation. The initiation of the initial segments starts during the final phase of the inhalation, which was almost synchronous with the onset of the constriction. During the inhalation phase, speakers showed a prominent lowering gesture of the lower lip. This mouth opening gesture could either be related to the inhalation or to speech preparation. Speakers initiated the oral gestures for the nasal /n/ later than for /t/, relative to the exhalation onset. This timing seems to be sensitive to the identity of the initial segment and indicates a close coordination between respiratory and articulatory actions. The current study aims to extend the investigation of Rasskazova et al. (2019) on the coordination of respiratory, acoustic and articulation events prior to the utterance. We extended the analysis to 11 speakers, five initial segments /t/, /n/, / ʃ /, /a/, /h/ as well compare the temporal organization for two types of silent pre-speech intervals: utterance initial – before the first utterance and inter-speech pause – between first and second utterance.

Methods. Respiration, speech kinematics and acoustics were simultaneously recorded by means of Electromagnetic Articulography (EMA AG501) and Inductance Plethysmography. Eleven native German speakers, aged between 22 and 38 years, participated in the study. The participants read utterances aloud, which were presented in randomized order on a computer screen. The speech material involved eleven utterances that consisted of two sentences each. They were mixed with various filler sentences, which differed in their structure and consisted of one sentence only. The target utterance was controlled for sentence length and word stress. The initial segment of the first word varied between initial /t/, /n/, /ʃ/, /a/, /h/, respectively. To compare the coordination of respiratory and articulatory events in utterance initial silent interval and inter-speech pause, the target sentences start with the same initial segment. Five repetitions of each target utterance were produced. All data were labelled with the visualization and labelling tool MVIEW (Tiede 2005), written in MATLAB. Temporal respiratory events were labelled as inhalation minima for the onset of inhalation and maxima for the onset of exhalation. The movement onset of the lower lip and the articulatory gestural phases of the targeted segment were determined automatically by using a 20% threshold criterion of the tangential velocity signal. For words starting with /h/ the dorsal gesture towards the following vowel was measured. Temporal coordination was investigated in two positions: prior to the utterance onset of each stimulus (henceforth: utterance-initial) and prior to the second sentence of the stimuli (henceforth: inter-speech pause). The following timepoints were included: acoustic onset of speech, inhalation and exhalation onset, movement onset of the lower lip as well as the tongue tip gesture of the upcoming segments. To normalize time, the exhalation onset was taken as a reference point and all events were subtracted from it.
Statistics for the presented data were carried out by calculating linear mixed effect models to test differences for latencies for different segments and sentence conditions as well as by calculating Relative Standard Deviation (RSD) to test the variability using R 4.2.1 (R Core Team 2022).

Results. Figure 1 shows that the inhalation in both utterance initial and inter-speech pause position starts as a first preparatory event, on average 760 ms in utterance initial condition (ui) and 496 ms in inter-speech pause (isp). The identity of the initial segment does not affect inhalation duration. The movement onset of the lower lip starts after inhalation with a delay of 484 ms in ui and 200 ms in isp. Unlike inhalation, lower lip latency does not differ significantly with sentence position. The acoustic onset, in many cases, starts after exhalation onset. The acoustic onset is initiated later in utterance initial condition (114 ms) compared to inter-speech (69 ms). This delay is significantly different for /ʃ/ ($t = 5.7$, $p < 0.001$): 60 ms for ui condition and only 24 ms for isp being almost synchronous with exhalation and nucleus onset phase of the oral gesture. The largest variation (RSD over 64%) in the delay between acoustic onset and exhalation were found for /n/

for both sentence conditions. For each initial segment, onset of the articulatory gesture starts prior to the exhalation and shows relatively high variability. The onset for vocalic gestures /a/ and /h/ starts earlier than for consonants. The gestural onset of /ʃ/ is the closest to the exhalation, most consistent in timing and stable across all speakers and sentence conditions. The articulatory gesture phases and exhalation onset are as expected tightly coupled with each other at the constriction phase. For all initial segments, the closest phase to exhalation onset is the onset of the constriction phase (nucon_gesture in Figure 1). For /t/ and /n/ the onset of the constriction is almost synchronous with exhalation ($t = 3.6$, $p < 0.001$), on average around 30 ms prior to the exhalation. In summary, we found that inhalation always starts first as a preparatory activity followed by the mouth opening gesture. There is no link between inhalation onset and the mouth opening gesture as well no effect of the initial segment on the timing of the mouth opening gesture. However, both identity of the initial segment as well sentence condition affects the timing of the acoustic onset and gestural phases of the initial segment. During utterance initial position the latencies are longer and show more variability. The initial segment /ʃ/ shows the least variability across phases for articulatory gesture in both sentence conditions.

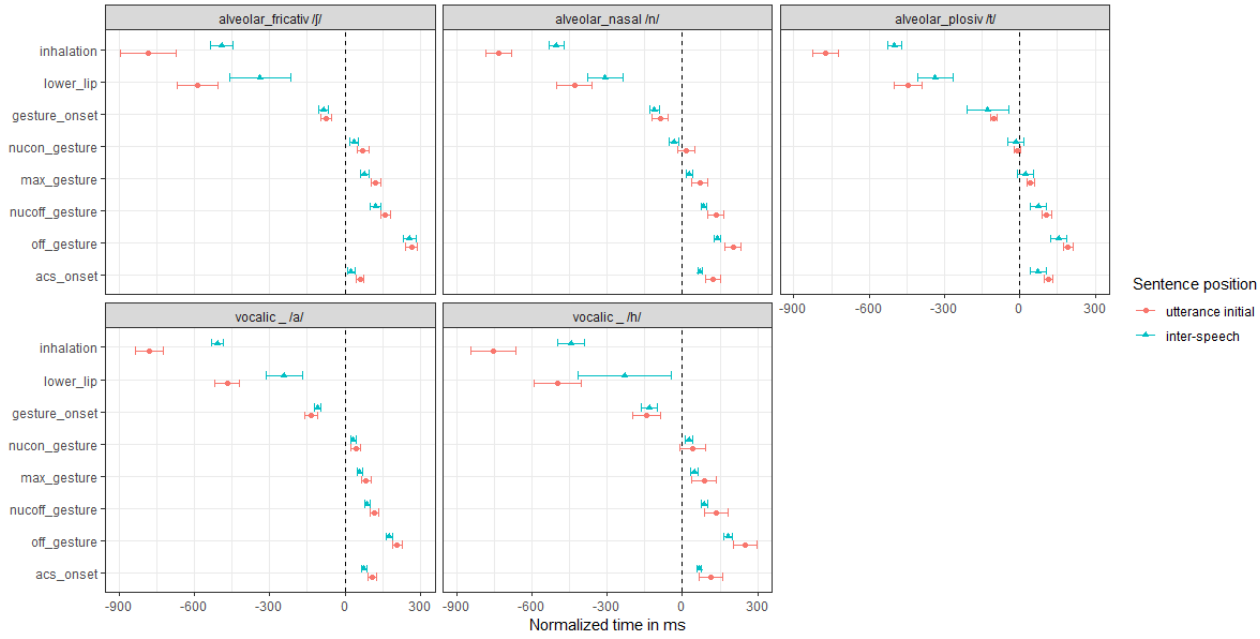


Figure 1: *Timing and the average duration of preparatory events prior to the speech initiation for each initial segment (subplots) during utterance initial silent interval (red) and inter-speech pause (blue).* On The y-axis are *preparatory events prior to speech. On the x-axis is normalized time in ms relative to the exhalation onset (0).*

Discussion. In this study we confirm our previous results (Rasskassova et al. 2019) on temporal alignment between oral articulators and onset of exhalation. These latencies showed relatively low variability and show that the initiation of articulatory gestures starts during the final phase of the inhalation, which can be interpreted as evidence for a close coupling between respiratory and oral actions for gestural organization. The timing of gestural organization and the acoustic onset of initial segment is sensitive the identity of initial segment and the sentence position. During inter-speech pauses the latencies are generally shorter, especially for the inhalation duration, indicating that speakers vary their speech preparatory activities according to the sentence conditions. Although, in the inter-speech sentence condition there was less variability than in utterance initial condition, it was still surprisingly high for some cases. The presented data summarizes the results across speakers. The interindividual variation may shed another light on these findings.

References

Fuchs, S., Petrone C., Krivokapić, J. & Hoole P. (2013). Acoustic and respiratory evidence for utterance planning in German. Journal of Phonetics, 41 (1), 29–47.

Mooshammer, C., Goldstein, L., Nam, H., McClure, S., Saltzman, E. & Tiede, M (2012). Bridging planning and execution: Temporal planning of syllables. Journal of phonetics, 40 (3), 374–389.

R Core Team (2022). R: A language and environment for statistical computing, Wien.

Tiede, M. (2005). Mview: software for visualization and analysis of concurrently recorded movement data, New Haven, CT: Haskins Laboratories.

Rasskazova, O., Mooshammer, C. & Fuchs, S. (2019). Temporal coordination of articulatory and respiratory events prior to speech initiation. Proceedings of the Conference Interspeech, 884-888.

Dialect specific patterns of gestural timing? Evidence from lateral clusters.

Emily Gorman[1]

[1]*Lancaster University*

e.gorman@lancaster.ac.uk

Introduction. Across the world's languages, there are a small number of timing topologies that govern the temporal structures of speech. For example, a C-Centre pattern is considered the typical pattern for onset clusters in branching languages (Browman and Goldstein, 2000). This pattern predicts a constant relationship between the centre of the onset consonant cluster and the vowel, regardless of cluster complexity. From the perspective of Articulatory Phonology, (Browman and Goldstein, 1988), this phenomena is explained by the competitive coupling of the pre-vocalic consonant gestures (or more specifically, of the planning oscillators associated with the gestures) which are coupled anti-phase to one another, but in-phase to the vowel. This results in a pattern of symmetrical shifting of onset consonants towards and away from the vowel (Browman and Goldstein, 1988). One important question is: to what extent do such timing patterns interact with the phonetic quality of the participating segments? This issue is particularly prevalent in the case of laterals where inconsistent timing patterns have been observed. For example, German has clear onset laterals, which show an asymmetric shift pattern in onset clusters (Mücke, Hermes, and Tilsen, 2020), whereas American English has dark onset laterals, which show a C-Centre pattern (Marin and Pouplier, 2010). Lateral darkness thus appears to play a mediating role in the cross-linguistic differences in onset lateral timing. For this reason, onset lateral clusters provide an interesting testing case for illuminating the role of syllable structure in conditioning consonant timing. The effect of lateral darkness is, however, potentially confounded by language differences. This is resolved here through a within-language approach, which allows the temporal effects of phonetic variation in /l/ to be isolated. Two dialect regions of British English, Standard Southern British English (SSBE), and Lancashire/ Manchester English, are compared. SSBE is observed to have clear /l/ onsets and dark /l/ or vocalised codas (Turton, 2014), while Lancashire/Manchester has dark /l/s in all positions (Hughes, Trudgill, and Watt, 2012). A further contribution of this study is the comparison of /l/ onset clusters across multiple vocalic contexts (e.g., see Strycharczuk, Derrick, and Shaw (2020) for the effect of vowel on /l/.)

Methods. Data was collected using audio-synchronised electromagnetic articulography (Carsten AG501, DPA4006A microphone). 14 participants were recorded; 8 were self-reported speakers of SSBE, and 6 of Lancashire or Manchester English. Sensors were glued mid-sagittally to the tongue tip, tongue body, tongue dorsum, lower gum line, and upper and lower lips. Participants read sentences containing /l/ in complex and singleton onset pairs, e.g., "Say tea lug again" (singleton), and "Say tea plug again" (cluster). Post-lateral vowels varied between two contexts, " l**i**p", and "l**u**g", and consonant clustered with /l/ varied between /k/ and /p/. Each unique vowel-consonant combination was repeated 4 times. The time of gestural target achievement was identified; this was defined as the point in time when the relevant sensor, in the relevant dimension, achieved its velocity minimum. For example, for /l/, the relevant sensor was the tongue tip in the vertical dimension. Lag measures were taken between the lateral and anchor consonant targets and compared across singleton and cluster pairs (e.g., in the "plug" "lug" pair, /g/ was the anchor consonant). Since a C-Centre pattern requires a rightward shift of C2, in this case /l/, towards the vowel, a C-Centre timing pattern predicts that there should be a smaller lag for clusters than for singletons within each pair. Measures were also taken of cluster C-Centre lags, i.e., the time between the anchor target and a point equidistant between the velocity minima of C1 and C2. Cluster C-Centre lags were compared to singleton C-Centre lags to determine whether the C-Centre to anchor lag was stable across singleton and cluster pairs. Additional cluster intervals were measured and compared to the singleton C-Centre lag, namely the Left Edge lag (C1 to anchor) and the Right Edge lag (C2 to anchor), to further determine the stability of these intervals across the singleton and cluster pairs. Lags were first compared visually (see Figure 1 below), before statistical comparisons were made through anovas of linear mixed effects models using the *lme4* R package (Bates et al., 2015). A C-Centre timing pattern predicts a stable relationship between the onset C-Centre and anchor across singleton and cluster contexts, hence predicts least variability between C-Centre measures.

Results and Discussion. Compared across singleton and cluster pairs, the /l/ to anchor lag was consistently shorter in the cluster context for both dialects, consistent with a C-Centre timing pattern. A more direct measure of C-Centre stability, the comparative stability measures shown in Figure 1, however, found that across the singleton-cluster pairs, the C-Centre lag was just as stable as the Right-Edge lag for both dialects. While inconsistent with a C-Centre effect, this was not a particularly surprising finding given the similar inconsistent timing patterns found, for example, in Mücke, Hermes, and Tilsen (2020). A more surprising finding of the study was the lack of a qualitative difference between the relative timing patterns of the two dialects; this is noteworthy since there was good reason to expect a difference here. It may be the case that a more extreme difference in lateral darkness (such as the difference between American English and German lateral onsets) is required to produce qualitative differences in cluster timing patterns. Alternatively, these results may suggest that it is not the phonetic properties of /l/, but rather its phonological status, which trigger the differences in onset later timing found in previous studies.

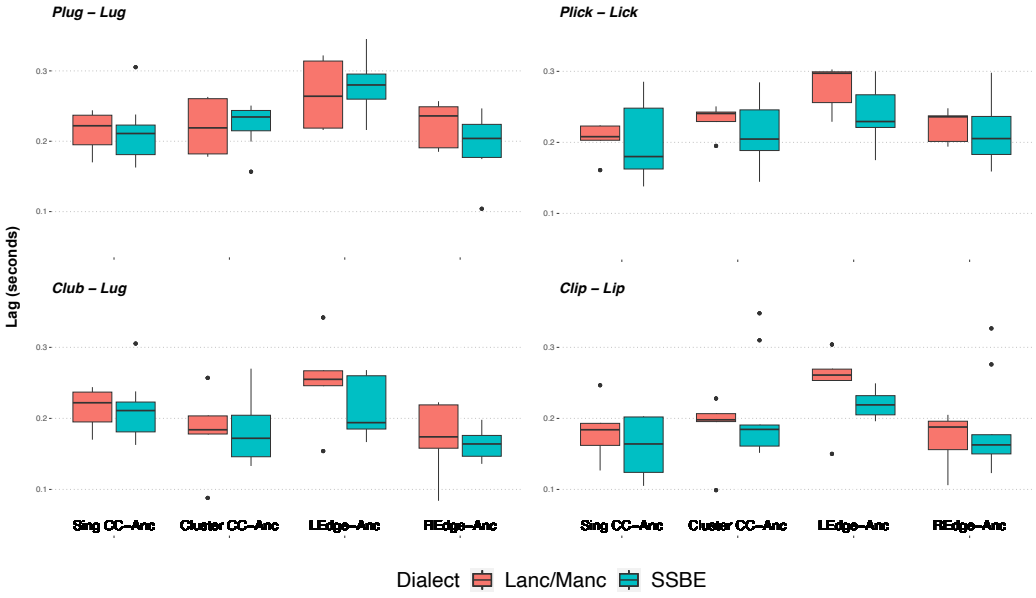


Figure 1: *Stability lags for vowel/cluster combinations: plug (top left), plick (top right), club (bottom left), clip (bottom right). Colour indicates dialect. The further away the cluster lag relative to that of the Singleton CC lag, the less stable it is across the singleton-cluster pair.*

References.

Bates, Douglas, Martin Maechler, Ben Bolker, and Steve Walker (2015). "Fitting linear mixed-effects models using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48.

Browman, Catherine P and Louis Goldstein (1988). "Some notes on syllable structure in articulatory phonology". In: *Phonetica* 45.2-4, pp. 140–155.

Browman, Catherine P. and Louis Goldstein (2000). "Competing constraints on intergestural coordination and selforganization of phonological structures." In: *Bulletin de la Communication Parlée* 5, pp. 25–34.

Hughes, Arthur, Peter Trudgill, and Dominic Watt, eds. (2012). *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles*. Fifth. London: Hodder.

Marin, Stefania and Marianne Pouplier (2010). "Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling odel". In: *Motor Control* 14.3, pp. 380–407.

Mücke, Doris, Anne Hermes, and Sam Tilsen (2020). "Incongruencies between phonological theory and phonetic measurement". In: *Phonology* 37.1, pp. 133–170.

Strycharczuk, Patrycja, Donald Derrick, and Jason Shaw (2020). "Locating de-lateralization in the pathway of sound changes affecting coda/l". In.

Turton, Danielle (2014). "Some /l/s are darker than others: accounting for variation in English /l/ with ultrasound tongue imaging". In: *University of Pennsylvania Working Papers in Linguistics* 20.2, pp. 189–198.

Phase-locking of Articulation and Spectral Flow

Jessica Campbell[1], Louis Goldstein[1], Leonardo Lancia,[2,3]

[1]Department of Linguistics, University of Southern California
[2]Laboratorie Parole et Langage, Aix-Marseille Université / CNRS
[3]Laboratoire de Phonétique et Phonologie, Université Sorbonne Nouvelle / CNRS
jac95339@usc.edu, louisgol@usc.edu, leonardo.lancia@cnrs.fr

Introduction. Sensorimotor integration of speech allows for the coordination of a speaker's speech production actions with the resulting speech sound. In the brain, this integration has been theorized to stem from coupling of neural oscillations in the auditory and motor cortices (Assaneo & Poeppel 2018), and it has been observed in their data on the repetition of discrete syllables at different rates. In addition, however, these phase-locked neural oscillations may be supported for continuous speech through phase-locking between the acoustic and articulatory signals themselves. This idea has previously been proposed by Goldstein (2019), who theorized that binding of sensory and motor signals during sensorimotor integration stemmed from the matching of articulation and acoustics in the time domain. Goldstein (2019) tested this by comparing the correlation of articulatory and acoustic modulation functions at different lags. In the current study, we approach this comparison by examining phase-locking between the signals, rather than their correlation, allowing behavior studies to be interpreted more cohesively with neural studies. Rather than comparing the signals directly, we can exploit the oscillatory nature of these signals by comparing the phases of their cycles. Such an analysis eliminates the direct comparison of amplitude (one cycle can maximally reach an amplitude of 2π). Thus, the current study directly examines the stability of the temporal relationship between acoustic and articulatory signals, employing an analysis of phase-locking values. Phase Locking Value (PLV) was determined by the variability of the results of a comparison between signals' phases at each moment in time (Lancia 2023). For example, two sine waves of the same frequency, one delayed by a phase of π, have a PLV of 1; the difference in phase between the signals has no variability.
To determine phase locking between the articulatory and acoustic domains, two low-dimensional signals were derived from the speech stream; both have been theorized to be "neurally viable" in that their frequencies occur in or near a range associated with high sensitivity of the auditory cortex and at which synchronization of auditory and motor cortices occurs, and in that they exhibit relatively low variability (Poeppel & Assaneo 2020, Campbell et all. 2023). For the current study, the acoustic signal examined was *spectral flow* (or the "acoustic modulation function" in Goldstein [2019]); this signal profiles the instantaneous change of MFCC parameters over time. The articulatory signal in the current study was the *articulatory modulation* function, which similarly profiles instantaneous change in the vocal tract articulators over time. The current study calculates this function from only the supralaryngeal articulation (excluding the Velum) available in the X-Ray Microbeam Corpus (Westbury et al. 1994). Given that spectral change is largely caused by supralaryngeal articulation, we hypothesize that the two signals will be significantly phase locked, despite previous findings of higher frequency for spectral flow than articulatory modulation (Goldstein 2019, Campbell et al. 2023). This would allow for the binding of the two signals during production and perception, and, assuming that neural activity does respond to these signals, would support the phase-locking of neural signals in the auditory and motor cortices during speech processes.

Methods. Data consisted of simultaneously collected articulatory pellet-tracking data and acoustic recordings in the X-Ray Microbeam Corpus from nine participants reading the "Grandfather Passage" (Darley et al. 1975). Spectral flow was calculated following Goldstein (2019) and Campbell et al. (2023) as the sum of the squared change in the second through thirteenth MFCC parameters across successive time samples. Articulatory modulation was calculated similarly as the sum of the squared Euclidean distance between seven marker positions across successive samples. Both were filtered with a ninth order 12Hz lowpass Butterworth filter. Instantaneous phase for each signal was estimated using a Hilbert transform (Lancia 2023). To determine PLVs, the signals for each speaker were windowed into 50 frame (~343ms) overlapping sections with a step size of five frames, an average of 652 windows per speaker. The mean difference in relative phase was then determined across each window (we assumed that spectral flow would oscillate with at most double the average frequency of articulatory modulation). To determine whether phase locking only occurs because of the periodicity of the signals, a control signal for each speaker was established: the articulatory modulation function was separated into two halves, and the second half placed before the first. This control signal thus contained the same periodicity as the original, and was still derived from the same utterance, consisting almost entirely of continuous speech. A one-sided sign test compared the PLVs pairwise obtained from the spectral flow and articulatory modulation function ("test condition") and the spectral flow and control signal ("control condition") for all speakers and windows (5,866 total pairs). A finding of

significantly higher PLVs in the former comparison was predicted and would show that the phase-locking was due to the specific context of the utterance, not simply the similarly periodic nature of the two signals.

Results. The median test PLV across all speakers and windows was 0.62, and the median control PLV was 0.58 (Recall that the maximum Phase Locking Value is 1). The median change from a test PLV to a control PLV calculated from the same spectral flow window was -0.037. A sign test comparing PLVs for every window within every speaker's utterance revealed that this difference was significant ($p < 0.001$). As predicted, the phase-locking values are higher when the articulatory signal is directly derived from the same speech signal as the spectral flow.

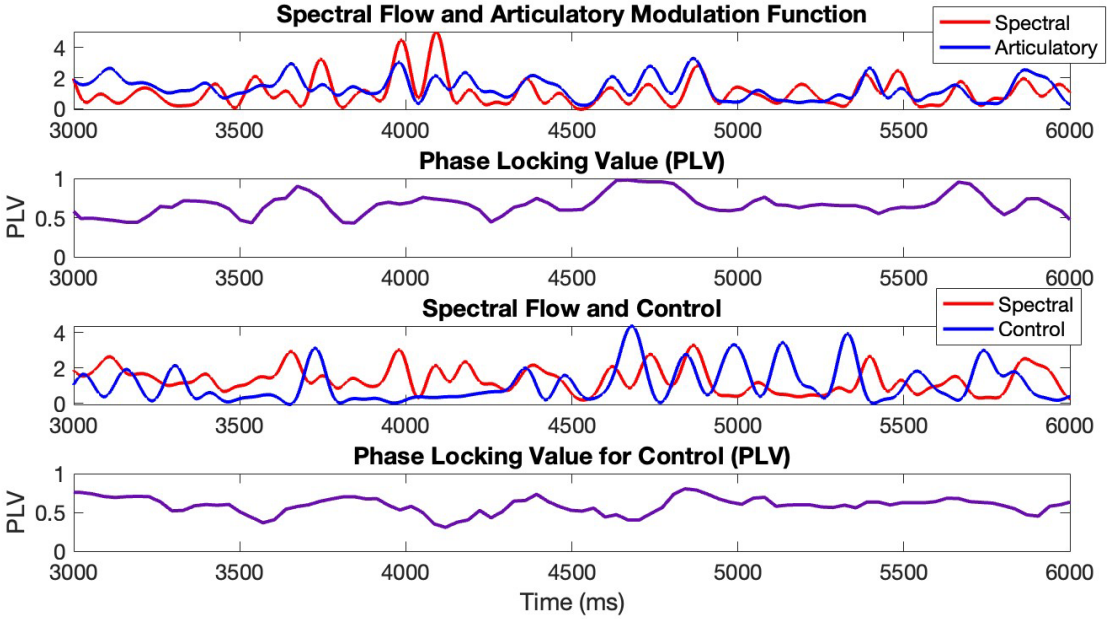


Figure 1: *The spectral flow, articulatory modulation function, and control signal for a portion of speech from one speaker, along with their PLVs. Spectral flow amplitude has been divided by 10.*

Summary and discussion. The results demonstrate a tendency toward phase-locking between spectral and articulatory information. Further research should examine whether the particular signals used in the study produce or are associated with time-locked neural activity, as this would more concretely connect the phase-locking observed in the current study with the phase-locking observed in neural ones. While the control did not produce as high PLVs as the test, it is notable that they were still fairly high. This likely stems from the high regularity of the signal, due both to the quasi-periodic nature of spectral flow and articulatory modulation and to smoothing introduced by the filtering process; creating a control version of a perfectly periodic wave the same way would produce extremely high PLVs. The finding of higher PLVs in the test condition may thus indicate that some *irregularity* in these conditions facilitates the binding of the two specific signals. This binding could then underwrite sensorimotor integration.

References.

Assaneo, M. F., & Poeppel, D. 2018. The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science Advances*, *4*(2), 1–9.

Campbell, J., Byrd, D., Goldstein, L. (August, 2023). Viable signal periodicities in speech rhythm. In Radek Skarnitzl & Jan Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences. Guarant International. 659-663.

Darley, F. L., Aronson, A. E., Brown, J. R. 1975. *Motor Speech Disorders*. W. B. Saunders Co.

Goldstein, L. 2019. The role of temporal modulation in sensorimotor interaction. *Frontiers in Psychology*, 10, 1–12.

Lancia, L. (2023). Instantaneous phase of rhythmic behaviour under volitional control. *bioRxiv*, 2023-11

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, *21*(6), 322-334.

Westbury, J. R., Turner, G., Dembowski, J. 1994. *X- ray Microbeam Speech Production Database User's Handbook*. University of Wisconsin.

Oral session 8

Phonetics/Phonology II

11:00– 12:00 am

	Title	Authors
11:00 - 11:20 am	Contrasting phonetic effects of morphological boundaries for vowel and consonant suffixes	Motoki Saito (University of Tübingen)*
11:20 - 11: 40 am	The role of tongue body position in obstruent-lateral cluster palatalization: Evidence from Spanish EMA data	Andrea García-Covelo (IPS-LMU Munich, IKER-UMR5478, UPPA)*; Marianne Pouplier (LMU); Ander Egurtzegi (CNRS-IKER)
11:40 - 12:00 am	A Real-time MRI Investigation of Larynx Raising in Amharic Ejectives	Lavinia Price (Institute for Phonetics and Speech Processing, LMU Munich)*; Marianne Pouplier (LMU); Philip A Hoole (Institute of Phonetics, Munich University)

Contrasting phonetic effects of morphological boundaries for vowel and consonant suffixes

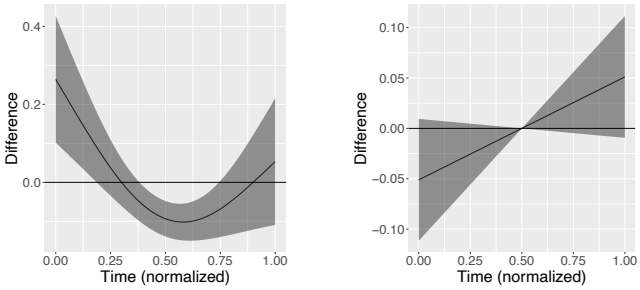
Motoki Saito[1]

[1] *University of Tübingen*
motoki.saito@uni-tuebingen.de

Introduction. Morphological structures are not available to determine phonetic realizations (e.g., Levelt, Roelofs, and Meyer, 1999). Challenging this assumption, a number of studies have found different phonetic realizations for different morphological structures (e.g., Plag, Homann, and Kunter, 2017). While it is getting clearer that morphological structures do affect phonetic realizations, it still remains unclear what phonetic differences should be expected for segments at a morphological boundary. Some studies have found longer duration and more peripheral tongue positions, namely phonetic enhancement (e.g., Seyfarth et al., 2017), while others have found shorter duration and more centralized tongue positions, namely phonetic reduction (e.g., Plag, Homann, and Kunter, 2017), for segments at a morphological boundary. One possible confounding factor is sonority of segments under investigation. Duration-lengthening may increase perceptability of vowels, while it may not be the best way to increase perceptability of consonants. In fact, reduction effects of a morphological boundary are often found for consonants (e.g., Plag, Homann, and Kunter, 2017). It has also been suggested that vowels and consonants are affected differently by a morphological boundary (Smith, Baker, and Hawkins, 2012). The current study, therefore, investigates the interaction between sonority and morphological-boundary effects.

Methods. All the German words with the word-final *-er* [ɐ] (highly sonorant) and *-t* [t] (less sonorant) were collected from the Karl-Eberhards Corpus of spontaneously spoken southern German (KEC) (Arnold and Tomaschek, 2016). These two word-final segments can be non-morphemic (e.g., *Vater* [faːtɐ] / *Luft* [lʊft]) as well as morphemic (e.g., *klein+er* [klaɪn+ɐ] / *sag+t* [zaːk+t]). Phonetic realizations of these word-final segments were analyzed in terms of their acoustic duration, mean tongue heights, and tongue trajectories, using Generalized Additive Mixed-effects Models (GAMMs) and Quantile Generalized Additive Mixed-effects Models (QGAMMs). These models all had variables of suffix distinctions (i.e., *-er* vs. *-t*), morphological status of the word-final *-er*/*-t* (i.e., `Morph`), word frequency, and speaker differences. In addition, speech rate and word duration were also considered for the duration model and the two tongue-position models respectively. The tongue-position models also had previous and next segments as additional random effects. The tongue-trajectory model also had a smooth term of normalized time to model tongue trajectories, and the model was set up in such a way that differences between the two morphological conditions were directly captured by a single smooth curve (i.e. a difference curve) and therefore tested directly.

Results. Duration was longer for the morphemic *-er*, compared to the non-morphemic *-er* ($\beta = 0.098$, $p < 0.001$). This effect of a morphological boundary was siginificantly attenuated for the suffix *-t* ($\beta = -0.116$, $p < 0.001$). Another GAMM of the same model structure with the suffix *-t* as the reference level confirmed that the presence of a morphological boundary significantly reduced duration of *-t* ($\beta = -0.018$, $p < 0.001$). Similarly to the duration model, the mean-tongue-height model showed that the morphemic *-er* was articulated more clearly than the non-morphemic *-er* ($\beta = 1.290$, $p \approx 0.016$). This hyper-articulation effect of `Morph` was significantly attenuated for the suffix *-t* ($\beta = -1.421$, $p \approx 0.014$). Another QGAMM of the same model structure with *-t* as the reference level confirmed that there was no such effect of a morphological boundary for *-t* ($\beta = -0.131$, $p \approx 0.561$). Finally, the tongue-trajectory model for *-er* indicated significant differences between the morphemic and non-morphemic *-er* (edf $= 1.942$, $p < 0.001$), while the model for *-t* predicted no difference between the morphemic and non-morphemic *-t* (edf $= 1.000$, $p \approx 0.098$). The articulation of the morphemic *-er* was predicted to begin at a higher position than the non-morphemic *-er* and get lowered significantly lower than the non-morphemic *-er* at the center of the vowel (Figure 1a). By contrast, there is no difference in tongue positions between the morphemic and non-morphemic *-t* throughout the segment, indicated by the entire confidence intervals containing the horizontal black line at y=0 in Figure 1b.



(a) The suffix *-er* condition (b) The suffix *-t* condition

Figure 1: Predicted differences in tongue trajectories between the morphemic and non-morphemic *-er* (left) and *-t* (right). No difference is predicted where confidence intervals contain the horizontal line (y=0).

Discussion. Longer duration for the morphemic *-er* and shorter duration for the morphemic *-t* indicate enhancement and reduction effects of a morphological boundary for vowels and consonants respectively. It is compatible with enhanced vowels and reduced consonants found in English prefixes by Smith, Baker, and Hawkins (2012). In addition, these current results resolve the seemingly-contradictory findings in the literature regarding longer and shorter duration induced by a morphological boundary (Plag and Ben Hedia, 2018; Plag, Homann, and Kunter, 2017). Following these observations of duration, clearer articulation was also observed for the morphemic, compared to the non-morphemic *-er*, while the morphemic and non-morphemic *-t* did not show significant differences in mean tongue heights. These results suggest that phonetic enhancement effects of a morphological boundary are limited to vowels. Clearer articulation of the morphemic *-er* was mainly observed at the center of the vowel. Higher tongue positions at the onset of the vowel also indicate clearer articulation, because more articulatory effort has to be paid for more dynamic tongue movements (e.g., Lindblom, 1983). Contrary to *-er*, the suffix *-t* did not show any difference in tongue trajectories between the morphemic and non-morphemic conditions. These results echo the duration and mean-tongue-height models and indicate that phonetic enhancement effects of a morphological boundary are limited to vowels. None of these results is predicted by a feed-forward modular-based model (e.g., Levelt, Roelofs, and Meyer, 1999). The current observations cannot be explained solely by phonological factors such as the number of syllables in the utterance/word, utterance-, word-, and segment-durations, or within-utterance positions. It was confirmed by a post-hoc analysis, in which the current observations were maintained after the statistical control of these additional phonological variables. The current results are rather compatible with a model that allows direct interactions between morphology and phonetics such as the Discriminative Lexicon Model (Baayen et al., 2019). These results, therefore, provide a small but important step forward regarding the necessity of revisiting the classical distinction of morphology, phonology, and phonetics, as well as practically helping to resolve seemingly-contradictory previous findings involving morphological boundary effects. For future research, a wider variety of segments from a wider range of sonority should be included for better generalizability.

References.

Arnold, Denis and Fabian Tomaschek (2016). "The Karl Eberhards Corpus of spontaneously spoken southern German in dialogues — audio and articulatory recordings". In: *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, pp. 9–11.

Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan, and James P. Blevins (2019). "The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning". In: *Complexity* 2019, pp. 1–39. DOI: 10.1155/2019/4895891.

Levelt, Willem J. M., Ardi Roelofs, and Antje S. Meyer (1999). "A theory of lexical access in speech production". In: *Behavioral and Brain Sciences* 22, pp. 1–75.

Lindblom, Björn (1983). "Economy of speech gestures". In: *The Production of Speech*. Ed. by Peter F. MacNeilage. New York: Springer-Verlag. Chap. 10, pp. 217–245.

Plag, Ingo and Sonia Ben Hedia (2018). "The phonetics of newly derived words: Testing the effect of morphological segmentability on affix duration". In: *Expanding the Lexicon: Linguistic Innovation, Morphological Productivity, and Ludicity*. Ed. by Sabine Arndt-Lappe, Angelika Braun, Claudine Moulin, and Esme Winter-Froemel. Berlin: De Gruyter, pp. 93–116. DOI: 10.1515/9783110501933-095.

Plag, Ingo, Julia Homann, and Gero Kunter (2017). "Homophony and morphology: The acoustics of word-final S in English". In: *Journal of Linguistics* 53.1, pp. 181–216. DOI: 10.1017/S0022226715000183.

Seyfarth, Scott, Marc Garellek, Gwendolyn Gillingham, Farrell Ackerman, and Robert Malouf (2017). "Acoustic differences in morphologically-distinct homophones". In: *Language, Cognition and Neuroscience* 33.1, pp. 32–49. DOI: 10.1080/23273798.2017.1359634.

Smith, Rachel, Rachel Baker, and Sarah Hawkins (2012). "Phonetic detail that distinguishes prefixed from pseudo-prefixed words". In: *Journal of Phonetics* 40.5, pp. 689–705. DOI: 10.1016/j.wocn.2012.04.002.

The role of tongue body position in obstruent-lateral cluster palatalization: evidence from Spanish EMA data

Andrea García-Covelo[134], Marianne Pouplier[1], Ander Egurtzegi[23]

[1]IPS-LMU Munich, [2]CNRS, [3]IKER-UMR5478, [4]UPPA

andrea.garcia@phonetik.uni-muenchen.de, pouplier@phonetik.uni-muenchen.de, ander.egurtzegi@iker.cnrs.fr

Introduction. Obstruent-lateral (OL) clusters diachronically often resulted in post-alveolar or palatal outcomes in Romance, e.g., Latin CLAVEM 'key' > Galician /ʧ/*ave*, Spanish /ʎ/*ave*, Italian /kj/*ave*. In most Romance varieties, OL clusters show straightforward diachronic patterns of regular palatalization, irrespective of the voicing and manner of C_1 - in a C_1C_2V syllable - and of the position of the cluster within the word. In contrast, only /pl, fl, kl/ regularly palatalized in Galician, Portuguese, and Spanish, while palatalization was rare for /gl/ and practically non-existent for /bl/ (Repetti & Tutle 1987; Zampaulo 2019). This suggests that the manner and voicing of C_1 played a role in this sound change. Also, the position of the cluster (García-Covelo, in prep.) and lexical stress (Wireback 1997) may have interacted with palatalization. The outcomes of OL palatalization in Ibero-Romance can partly be explained by competing sound changes which preceded (or coexisted with) the palatalization process and may have blocked it by changing its triggering context (Garcia-Covelo, in prep.). One example is lenition, which caused post-vocalic voiced stops to be spirantized or deleted. Consistent with this hypothesis is that in contexts where voiced stop lenition was likely to occur, e.g., post-vocalically and word-initially (not utterance-initial), no palatalization is observed. It is generally agreed that OL palatalization must have had an articulatory origin, from an overlapping production of C_1 and C_2 in /kl, gl/ (Recasens 2018). However, the exact mechanisms behind this palatalization process through articulatory overlap are unclear. In this scenario, the palatalization of /l/, which is thought to be the first step in OL palatalization, would originally occur due to articulatory blending only when C_1 was a dorsal stop, not in the case of labial consonants. Yet, in the case of a voiced dorsal stop in those contexts where lenition was likely to occur, spirantization of the dorsal would potentially block the palatalization of /l/. The aim of this study is to test these assumptions and assess: 1. Whether articulatory dynamics may have indeed triggered OL palatalization; and 2. Whether the voicing of C_1, the position of the cluster within a word, and lexical stress affect these dynamics.

Methods. Articulography data from 10 speakers of Peninsular Spanish was acquired with synchronized audio. The experiment consisted of a reading task with the target stimuli embedded in the carrier phrase *Ahora diga X, por favor* 'Now say X, please'. For the current analysis, stimuli containing /kl, gl, l/ in three positions within the word (word-initial, post-consonantal and post-vocalic) and with three lexical stress patterns (stressed, pretonic and posttonic) were analyzed (/l/ only in stressed syllables). Each token was repeated five times in a randomized order. The articulatory landmark segmentation was performed for /l/, using the tangential velocity of the tongue tip sensor, to identify the constriction plateau, i.e., the points of constriction formation, maximum constriction, and constriction release (cf. Pouplier et al. 2022). The point of maximum constriction was used as the timepoint to extract the position of the sensor placed posterior to the tongue blade (henceforth: tongue body), which was used to study the palatalization of /l/ (cf. Kochetov 2005). To evaluate whether the coarticulatory changes in /l/ conditioned by a preceding dorsal stop would be in the direction expected for palatalization, i.e., whether tongue body during /l/ would be higher and/or more anterior in clusters compared to singleton, two linear mixed-effects models were fit with horizontal/vertical sensor position as response and phone (cluster/singleton) in interaction with position within the word as predictors. To assess whether these coarticulatory changes in /l/ are affected by lenition or lexical stress, i.e., whether tongue body during /l/ would be higher and/or more anterior in /kl/ compared to /gl/, two additional linear mixed-effects models were fit with horizontal/vertical sensor position as response and phone (/kl, gl/) in interaction with position within the word and with lexical stress as non-interacting variable as predictors. As the predictors are categorical variables with more than two levels, pairwise post-hoc t-tests (R package *emmeans*) will be reported for significant main effects ($\alpha = 0.001$ due to Tukey p-value adjustment).

Results. Data visualization suggested that tongue body during /l/ at maximum constriction is higher in /kl, gl/ than in /l/. No differences in the anteriority of the tongue body during /l/ were discerned. The models confirmed these observations (Table 1). Further visualizations showed that tongue body during /l/ is higher in /kl/ than in /gl/ in lenition-inducing positions, i.e., word-initially, and post-vocalically (Figure 1), but not post-consonantally, where the height values are similar. No effects of lexical stress could be discerned. Similarly, a faint tendency for /l/ to be slightly more anterior in /kl/ than in /gl/ was observed but no effect of position of the cluster or of lexical stress was obvious. The models mostly

confirmed these observations, reporting a significantly higher tongue body in /kl/ than in /gl/, but only post-vocalically (Table 1). Figure 1 shows /gl, kl/ coded for voiced stop lenition, i.e., lack of visible or audible burst or of velar constriction: while /gl/ always lenites post-vocalically, word-initial lenition may depend on the presence of a prosodic boundary before the token. The tongue body during /l/ is lower if /gl/ is lenited but, if there is no lenition, the tongue body height values of /gl/ pattern with those of /kl/. As lenition was not a predictor, word-initial variation was probably identified by the model as inter-speaker variation, thus deeming the differences non-significant.

Table 1. Result summary of the four linear-mixed models

Comparisons	Significant effects	Non-significant effects
/l/ vs. /kl gl/	- higher tongue body in /kl, gl/	- more fronted tongue body in /kl, gl/
/kl/ vs. /gl/	- higher tongue body in /kl/ (post-vocalically) - more fronted tongue body in /kl/ (post-vocalically and post-consonantally)	- effect of stress in tongue body height and anteriority during /l/

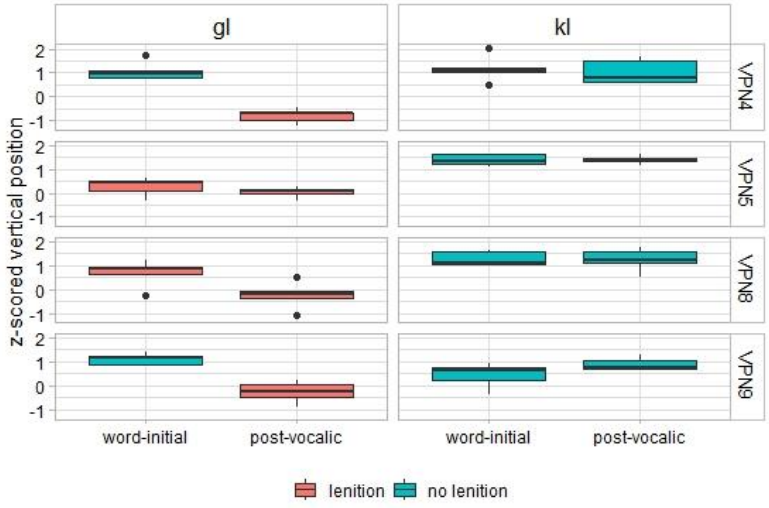


Figure 1. Z-scored values of /kl, gl/ in word-initial and post-vocalic position (four speakers). Higher values mean a higher sensor position.

Discussion. Articulatory dynamics seem to play a role in OL palatalization. During /l/, coarticulation with a preceding dorsal stop partly goes in the direction expected for palatalization, as we have a clearly higher but not fronter tongue body position in /kl, gl/ than in /l/. In addition, coarticulation during /l/ in /kl, gl/ is affected by voiced stop lenition but not by lexical stress; lenition decreases the amount of constriction, so that the tongue body is higher in /kl/ than in /gl/ in lenition-inducing contexts, i.e., post-vocalically and word-initially. The tongue body during /l/ is more anterior in /kl/ than in /gl/ post-vocalically and post-consonantally. These findings support the proposal that lenition may have played a role in the distribution of OL palatalization in Ibero-Romance. However, the acoustic and perceptual implications of the observed patterns remain to be explored to understand the possible coarticulatory origin of this sound change.

References

García-Covelo, A. (in preparation). *The development of obstruent plus lateral clusters in Ibero-Romance: a historical-phonetic approach to cluster palatalization.* PhD thesis.

Kochetov, A. (2005). "Phonetic sources of phonological asymmetries: Russian laterals and rhotics". *Proceedings of the 2005 annual conference of the Canadian Linguistic Association.*

Pouplier, M., Pastätter, M., Hoole, P., Marin, S., Chitoran, I., Lentz, T.O., Kochetov, A. (2022). "Language and cluster-specific effects in the timing of onset consonant sequences in seven languages." *Journal of Phonetics*, 93. 10115-3.

Recasens, D. (2018). *The production of consonant clusters. Implications for phonology and sound change*. Berlin/Boston: De Gruyter.

Repetti, L. & Tuttle, E. (1987). "The Evolution of Latin pl, bl, fl and cl, gl in Western Romance". *Studi Mediolatini e Volgari,* 33. 53–115.

Wireback, K. (1997). *The role of phonological structure in sound change from Latin to Spanish and Portuguese*. New York: Peter Lang.

Zampaulo, A. (2019). *Palatal sound change in the Romance languages. Diachronic and synchronic perspectives*. Oxford: Oxford University Press.

A Real-time MRI Investigation of Larynx Raising in Amharic Ejectives

Lavinia Price[1], Marianne Pouplier[1], Philip Hoole[1]

[1]Institute for Phonetics and Speech Processing (IPS), LMU Munich
{l.price|pouplier|hoole}@phonetik.uni-muenchen.de

Introduction. Larynx raising is the defining characteristic of ejective sounds (Catford 1977). Recent studies on ejectives across different languages, however, have cast doubt on this established view, calling into question the effectiveness or even necessity for the larynx to raise in order to produce ejectives (Brandt and Simpson 2021; Kingston 1985; Wright *et al.* 2002). Instead, supra-laryngeal cavity reduction is suggested as the key factor which may be achieved by various other articulatory strategies. As a consequence, our current understanding of ejective production by means of glottalic initiation is now being re-examined. The methodological difficulty of capturing larynx movement and pharyngeal volume represents an additional complicating factor in the debate surrounding ejectives. The current study seeks to contribute to this debate on the basis of Amharic, a Semitic language of Ethiopia, maximizing the advantages of real-time Magnetic Resonance Imaging (rt-MRI) to observe larynx raising along with supraglottal articulations. Gaining a better understanding of how laryngeal, supra-laryngeal, and aerodynamic factors coordinate in ejective production has important implications for the way in which we define control parameters in speech production and for our classification of non-pulmonic speech sounds.
Traditionally, ejectives are defined as products of larynx raising with simultaneous glottal and oral constrictions (Catford 1977). Larynx-raising reduces the supraglottal cavity, thereby elevating the intraoral air pressure (IOP). This results in the auditorily distinct quality of ejective release bursts. However, it has repeatedly been argued that larynx raising alone is insufficiently effective at increasing IOP to justify the intense bursts characterizing some ejectives (Kingston 1985). It has even been conjectured that ejectives can be produced without any significant laryngeal involvement at all (Brandt and Simpson 2021; Simpson 2014) with other factors conditioning the build-up of IOP. The nature of these other factors remains unclear, though. Kingston, based on Tigrinya, which is closely related to Amharic (Leslau 1997), proposed that supra-laryngeal articulations such as tongue root retraction and stiffening of the vocal tract walls may potentially be as important as larynx raising for supraglottal volume reduction (Kingston 1985). Larynx raising would then have a synergistic, rather than a defining role in ejective production and may variably be present in any given token. The high intra- and inter-language variation found in studies on the acoustic properties of ejectives, supports this view, suggesting that the realization of ejectives may be more variable than their textbook description allows for (e.g., Warner 1996; Wright *et al.* 2002). The rt-MRI analysis of ejectives in Amharic presents a unique chance to explore the complexities of ejective production.

Methods. We recorded rt-MRI data from eleven native speakers acquired at 50 frames per second. The stimuli included the voiceless pulmonic and ejective plosives of Amharic at three places of articulation (/p, t, k, p', t', k'/), as singletons and geminates (e.g., /t' vs. tt'/) in three word-positions (initial, medial, final) across 271 lexical items embedded within a carrier phrase. From the MRI images the vertical position of the larynx's lower edge was extracted via an edge-detection method, at onset (P1) and release (P2) of the target consonant's closure (Figure 1a, b). Larynx movement was then calculated as Δlarynx = P2 - P1. An acoustic analysis compared ejectives and their pulmonic counterparts in a number of measures known to characterize ejectives. These measures are burst intensity, burst duration and intensity of the post-burst voicing lag, which measures intensity over the interval following the burst release until onset of voicing. This interval is typically characterized by glottal frication (VOT) in pulmonic consonants but glottal closure in ejectives.

Results. Results from the ten speakers analyzed so far reveal that the larynx raises consistently and significantly during closure of Amharic ejectives at all three places of articulation (Figure 1c). A clear pattern is further shown at the acoustic level, where compared with their pulmonic counterparts the alveolar and velar ejectives are characterized by longer, more intense bursts (bilabials were excluded from the acoustic analysis). Velar ejectives are further characterized by an interval of glottal closure before voice onset of the following vowel.

Discussion. Importantly, our results support Catford's account by reaffirming that larynx raising is, at least for some languages, a defining characteristic of ejective production, despite recent contrary claims (Brandt and Simpson 2023; Simpson 2014; Sulaberidze *et al.* 2023). Nevertheless, we recognize that this description may not fully capture the ejective variability reported across different languages, which is influenced by linguistic, articulatory contexts, and individual speakers (Brandt and Simpson 2021; Kingston 1985). Thus, we will extend our analyses of laryngeal height, adding area function measures, which allow us to identify the possible role of supra-laryngeal articulators for cavity reduction

(Kingston 1985). By exploring the potential synergistic relationship in this way, we aim to uncover how laryngeal and supra-laryngeal articulations collectively influence ejective production. This analysis will shed more light on the role of cavity reduction in the production process, paving the way for insights into ejective realizations across languages. Such a global-gesture-approach (Mattingly 1990) to ejective production would enable us to consider the complexities of ejective sounds manifested cross-linguistically, while maintaining the Catfordian description, with the advantage that the presence of larynx raising would not represent the sole factor determining all ejective sound realizations. Furthermore, the result would have important implications for the way in which articulatory targets should be specified in speech production.

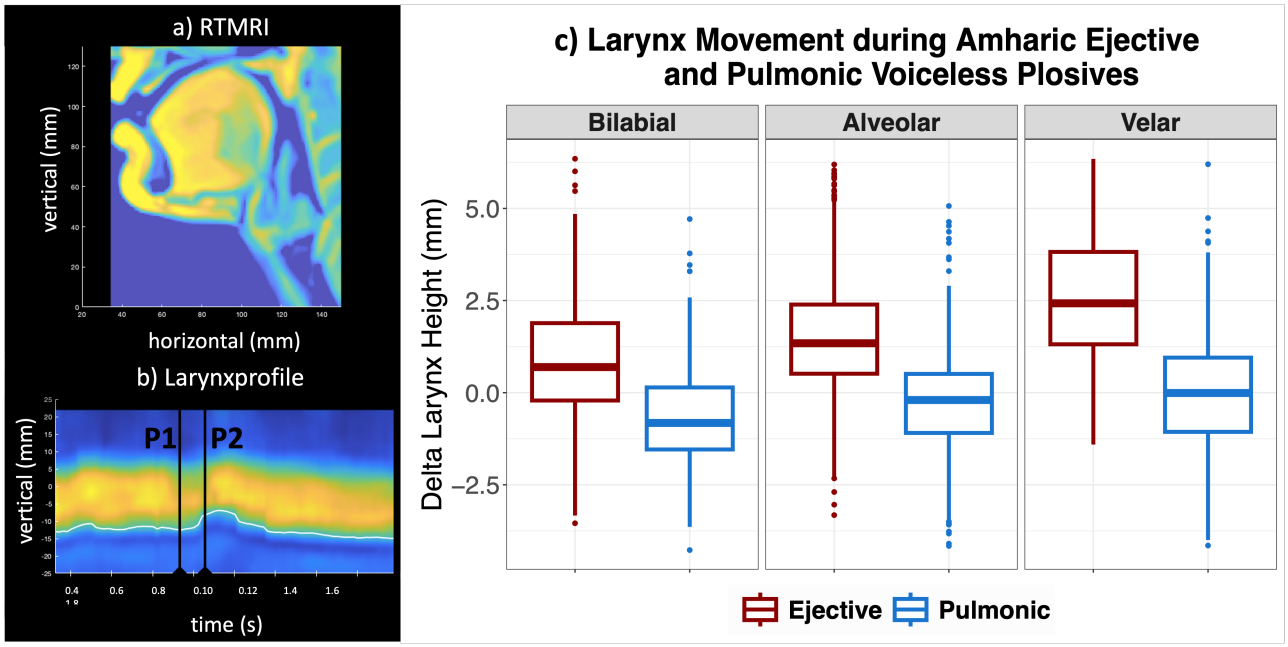


Figure 1: ***a)*** *Vocal tract configuration at the moment of release of the target consonant's closure, here an initial velar ejective.* ***b)*** *Intensity profile of the larynx for the entire Amharic sentence. The solid white line traces the larynx's lower edge. The positional values extracted at closure onset (P1) and release (P2) are indicated by the vertical black lines.* ***c)*** *Comparison of the amount of larynx movement (calculated as delta larynx = P2 – P1) during ejective and pulmonic voiceless plosives in Amharic by place of articulation.*

References

Brandt, E., & Simpson, A. P. (2021). The production of ejectives in German and Georgian. *Journal of Phonetics*, *89*, 101111.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Edinburgh: Edinburgh University Press.

Kingston, J. C. (1985). *The phonetics and phonology of the timing of oral and glottal events.* Ph.D. thesis, University of California, Berkeley.

Leslau, W. (1997). Amharic phonology. *Phonologies of Asia and Africa (including the Caucasus)*, *1*, 399-430.

Lindau, M. (1984). Phonetic differences in glottalic consonants. *Journal of Phonetics*, *12*(2), 147-155.

Mattingly, I. G. (1990). The global character of phonetic gestures. *Journal of Phonetics*, *18*(3), 445-452.

McGowan, R. S., & Saltzman, E. L. (1995). Incorporating aerodynamic and laryngeal components into task dynamics. *Journal of Phonetics*, *23*(1-2), 255-269.

Simpson, A. (2014). Ejectives in English and German. *Advances in Sociophonetics, John Benjamins*, 189-204.

Sulaberidze, N., Brandt, E., Hoole, P., Krämer, M., Reichenbach, J. R., & Simpson, A. P. (2023). Ejectives in Georgian. A real-time MRI analysis of vertical larynx movement. In: Radek Skarnitzl & Jan Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences (pp. 952–956). Guarant International.

Warner, N. (1996). Acoustic characteristics of ejectives in Ingush. In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96 (Vol. 3, pp. 1525-1528). IEEE.

Wright, R., Hargus, S., & Davis, K. (2002). On the categorization of ejectives: data from Witsuwit'en. *Journal of the International Phonetic Association*, 32(1), 43-77.

Oral session 9

Development

12:00 am– 01:00 pm

	Title	Authors
12:00 - 12:20 am	The effect of stress and word length on vowel reduction by Greek-speaking preadolescents and early adolescents	Polychronia Christodoulidou (Aristotle University of Thessaloniki)*; Katerina Nicolaidis (Aristotle University of Thessaloniki); Dimitrios Stamovlasis (Aristotle University of Thessaloniki)
12:20 - 12:40 am	The Hourglass of Speech: Modeling the Production/Perception Link	D. H. Whalen (CUNY/Yale)*
12:40 am - 1:00 pm	Age effect on intra-syllabic anticipatory labialization.	Louise Wohmann-Bruzzo (Université Sorbonne nouvelle)*; Nicolas Audibert (Laboratoire de Phonétique et Phonologie); Cecile FOUGERON (LPP)

The effect of stress and word length on vowel reduction by Greek-speaking preadolescents and early adolescents

Polychronia Christodoulidou, Katerina Nicolaidis, Dimitrios Stamovlasis

Aristotle University of Thessaloniki

polychri@enl.auth.gr, knicol@enl.auth.gr, stadi@edlit.auth.gr

Introduction. This study examines vowel reduction as a function of stress, word length and age in Greek. Vowel reduction occurs when vowels fail to reach their ideal target in the acoustic vowel space due to reduced vowel duration, spectral differences among adjacent speech sounds, and formant change rate (Moon & Lindblom, 1994). In some languages (e.g., English: Delattre, 1966), this phenomenon takes phonological extensions, resulting in the neutralization of reduced vowels, while, in other languages (e.g., Greek: Fourakis et al., 1999; Nicolaidis, 2003; Baltazani, 2007), the manifestation is phonetic, with reduced vowels occupying significantly more central positions in the acoustic vowel space, without altering their phonological quality. Generally, vowel reduction can occur due to various parameters, including stress, word length, focus, and speaking style (Moon & Lindblom, 1994; Fourakis et al., 1999; Nicolaidis, 2003; Baltazani, 2007). Regarding stress and word length, which are the factors investigated in the current study, Baltazani (2007) reported vowel centralization in the acoustic vowel space in unstressed conditions and in longer words for Greek-speaking adults due to reduced vowel duration in these conditions. Such findings align with Moon and Lindblom (1994), suggesting that as vowel duration decreases, the vowel undergoes a shift in the acoustic vowel space. Similar results were reported for Greek-speaking children up to 7 years by Christodoulidou et al. (2023), although for the younger age groups (three- and five-year-olds) of this study, the degree of temporal vowel reduction differed from that of adults and especially for three-year-olds, the correlation between normalized vowel space areas and relative vowel duration was not as strong as in adults. There is a scarcity of studies on vowel reduction in older age groups and for different languages although there is evidence that adult-like temporal and spectral vowel characteristics emerge around adolescence (Lee et al., 1999). Examining both temporal and spatial vowel reduction is particularly crucial, given that both duration and vowel space areas are linked to speech intelligibility (Metz et al., 1990; Sfakianaki et al., 2016). Such findings can be valuable for a comprehensive understanding of the developmental course of vowel reduction and spatiotemporal vowel organization, as well as for clinical intervention and advancements in speech technology.

Methods. Following up on our study (Christodoulidou et al., 2023), which examined vowel reduction in children up to 7 years of age, this cross-sectional experiment investigated 24 typically developing Greek-speaking preadolescents and early adolescents, evenly distributed across 3 gender-balanced age groups: nine-, eleven-, and thirteen-year-olds (i.e. including 8 participants, 4 males and 4 females, in each age group). Additionally, a control group of 8 adults (4 males and 4 females) was included. All participants engaged in a delayed repetition task, producing 15 Greek words of the form CV.CV.(CV) within the carrier phrase [ˈlɛo to ____ pɐˈdu] ‘I say ____ everywhere’. Each of the five Greek vowels, [ɐ, ɛ, i, o, u], was studied in the first syllable of triplets of phonetically similar two- and three-syllable target words (e.g., [ˈxɐ̲li] ‘mess’ - [xɐ̲ˈli] ‘carpet’ - [xɐ̲ˈlici] ‘gravel’), in which we examined (a) stressed vowels in disyllabic words, and (b) unstressed vowels, in the adjacent pre-stressed position, in both disyllabic (unstressed-2) and trisyllabic words (unstressed-3). Each word was produced five times, resulting in the analysis of a total of 2,400 vowels (32 participants × 3 stress/length conditions × 5 vowels × 5 repetitions). Measurements included (a) relative vowel duration calculated based on the duration of the two-syllable part shared among the words in each triplet and (b) normalized vowel space areas using F1 and F2 formant frequencies, which were Lobanov-normalized and rescaled into Hertz-like values (32 participants × 3 stress/length conditions × 5 repetitions = 480 areas). These variables were analyzed using linear mixed-effects models ANOVA in R, with speaker as a random-effects factor and age, gender, stress/length, and vowel (for relative vowel duration only) as fixed-effects factors. In these models, only statistically significant variables were retained. To investigate the degree of vowel reduction across ages, we also examined the percentage change in relative vowel duration between the stress/length conditions per speaker for each vowel separately and the percentage change in normalized vowel space areas between the stress/length conditions per speaker for each repetition separately using Wilcoxon tests (8 participants per age group × 5 vowels/repetitions = 40 values per age group). Finally, the relationship between relative vowel duration and normalized vowel space areas was also explored across ages with Pearson’s correlations (8 participants per age group × 3 stress/length conditions × 5 repetitions = 120 observations per age group).

Results. The results showed a decrease in relative vowel duration as age increased. Nine- and eleven-year-olds exhibited longer relative vowel duration than adults in the stressed condition, while no age-related differences were observed in the unstressed conditions. Gender showed no significant effect on this variable. With reference to the influence of stress/length, relative vowel duration decreased in the order stressed > unstressed-2 > unstressed-3 in each age group. The degree of vowel reduction between the stress/length conditions in adults was compared to that of the other age groups using Wilcoxon tests, and the results revealed only that eleven-year-olds exhibited significantly greater temporal vowel

reduction between the stressed and unstressed conditions compared to adults since relative vowel duration remained long until 11 years in the stressed condition and slightly decreased with age in the unstressed conditions. On the other hand, both age- and gender-related distinctions were absent in normalized vowel space areas. Regarding the influence of stress/length, normalized vowel space areas also decreased from the stressed to unstressed conditions due to the vowel centralization observed in the unstressed conditions, while statistically significant differences between the unstressed conditions in normalized vowel space areas were only observed in the overall participant group, not within age groups. Contrary to temporal vowel reduction, Wilcoxon tests showed no age-related differences in the degree of spatial vowel reduction between the stress/length conditions. In addition, despite the age-related variations in the degree of temporal vowel reduction, strong Pearson's correlations emerged between relative vowel duration and normalized vowel space areas at all ages, i.e. longer vowel durations were associated with larger vowel space areas ($r \geq 0.7$, $p < .0001$, see Figure 1). Figure 1, also, shows that higher values in both relative vowel duration and normalized vowel space areas were consistently present in the stressed condition compared to the unstressed ones across all age groups.

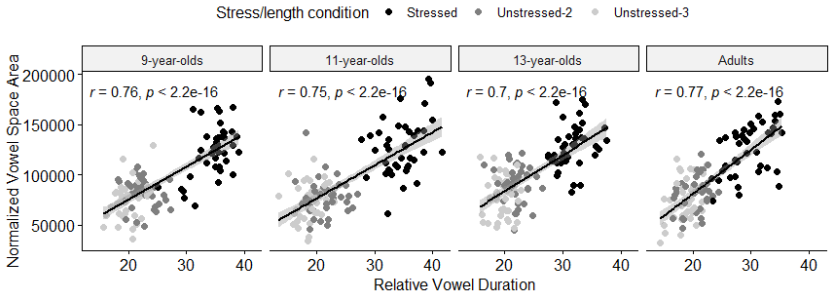


Figure 1: *Pearson's correlations between relative vowel duration and normalized vowel space areas by age.*

Discussion. Our findings showed that nine- and eleven-year-olds had significantly longer relative vowel duration than adults in the stressed condition, while no significant differences were observed in the unstressed conditions. Thus, differences emerged in how the stressed versus unstressed vowels were produced across ages with stressed vowels having longer duration till the age of 11 years and unstressed vowels declining in duration, which may be attributed to the less mature speech motor control in preadolescents. For Greek, the attainment of adult-like vowel reduction in the temporal domain seemed to occur at 13 years (cf. Kehoe et al. (1995) for English). With reference to spatial vowel reduction, no age-related differences were identified in normalized vowel space areas. Similarly to relative vowel duration, the stressed > unstressed-2 > unstressed-3 pattern was noted, but in this variable, the differences between the unstressed conditions were significant only in the overall dataset. Due to the absence of significant differences in normalized vowel space areas between the unstressed conditions across ages, our results are in partial agreement with Baltazani (2007) on the effect of word length. Finally, the correlation between relative vowel duration and normalized vowel frequencies, a relationship supported by Moon & Lindblom (1994) for adults, proved as robust for preadolescents and early adolescents as for adults. Therefore, spatiotemporal vowel organization does not appear to be affected by the age-related differences observed in vowel reduction, particularly in the temporal domain.

References

Baltazani, M. (2007). Prosodic rhythm and the status of vowel reduction in Greek. In *Selected Papers on Theoretical and Applied Linguistics from the 17th International Symposium on Theoretical & Applied Linguistics*, *14-17 April 2005* (Vol 1, pp. 31-43). Thessaloniki: Department of Theoretical and Applied Linguistics.

Christodoulidou, P., Nicolaidis, K. & Stamovlasis, D. (2023). Vowel reduction by Greek-speaking children: The effect of stress and word length. In *Proceedings of INTERSPEECH 2023*, *20-24 August 2023* (pp. 4773-4777). Dublin, Ireland.

Delattre, P. (1966). A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics in Language Teaching*, *Vol 4*, 183-198.

Fourakis, M., Botinis, A. & Katsaiti, M. (1999). Acoustic characteristics of Greek vowels. *Phonetica*, *Vol 56*, 28-43.

Kehoe, M., Stoel-Gammon, C. & Buder, E. (1995). Acoustic correlates of stress in young children's speech. *Journal of Speech and Hearing Research*, *Vol 38*, 338-350.

Lee, S., Potamianos, A. & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of Acoustical Society of America*, *Vol 105*, 1455-1468.

Metz, D.E., Samar, V.J., Schiavetti, N. & Sitler, R.W. (1990). Acoustic dimensions of hearing-impaired speakers' intelligibility: Segmental and suprasegmental characteristics. *Journal of Speech, Language, and Hearing Research*, *Vol 33* (No 3), 476-487.

Moon, S.J. & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of Acoustical Society of America*, *Vol 96* (No 1), 40-55.

Nicolaidis, K. (2003). Acoustic variability of vowels in Greek spontaneous speech. In M. Solé, D. Recasens & J. Romero (Eds.), *Proceedings of the XVth International Congress of Phonetic Sciences*, *3-9 August 2003* (pp. 3221-3224). Barcelona: Universidad Autónoma de Barcelona.

Sfakianaki, A., Nicolaidis, K. & Okalidou, A. (2016). Vowel production and intelligibility in hearing-impaired speech: Evidence from Greek. *Glossologia*, *Vol 24*, 75-92.

The Hourglass of Speech: Modeling the Production/Perception Link

D. H. Whalen[1]

[1]City University of New York; Yale Child Study Center
whalen@haskins.yale.edu

For decades, speech scientists have been debating the link between production and perception in speech. It is possible to study each separately, and many researchers do so, but if a speaker did not think that her speech would be perceived, she would not bother to speak. The processes that go into the speaker's choice of utterances to produce cover many layers of linguistics and cognitive science: Selecting a message to convey, associating the right words with the selection (taking aspects of the audience into account), getting the phonemes/gestures lined up and layering them with prosodic content, etc. An early conceptualization of this is the "chain of speech" model of Denes and Pinson (1963), which covers the multitude of levels without being very specific about the production/perception link. It is impossible to take account of everything: The physiological measures that are the immediate domain of ISSP are challenging enough for simple situations. Most of the complications must be ignored if we are to collect enough repetitions to overcome the noise in the system and our measurements of it. Ultimately, this amazingly complex process has to be squeezed through an acoustic system that drips out one pressure value at a time. (The visual and tactile signals contribute, at different time scales.) This drip can be adequately modeled as occurring once every 0.021 ms. That signal then goes to the listener, for an equally complex set of processes that makes sense of these pressure changes. Thus, the mechanics of production and perception differ so greatly that it seems odd that they have any connection at all. Indeed, the reason the debate has continued for so long is that the evidence for a lack of a link is intuitive and the existence of a link is hard to demonstrate (Pardo & Remez, 2021). My own thinking has lately taken shape around an analogy: Speech is like an hourglass.

Analogies are always limited in their application, of course, and taking an outmoded technology does not necessarily appear likely to be of use. Nonetheless, here is what appeals to me.

The upper part of an hourglass, while it is measuring time, is full of content (see Fig. 1). This is the speaker's domain, in which ideas are rich and interconnected (even if they shift during the course of an utterance). These ideas must be linearized if they are to enter the spoken realm. It is not possible to say everything at once, as might be the case with a visual display that remains available to the viewer indefinitely. Speech is evanescent, for better or worse. So these wonderful grains of ideas must become orderly and pass through the neck of the hourglass, one pressure value at a time, and enter the atmosphere below.

The lower half is the hearer's domain. She does not have direct access to all the wonderful ideas that the speaker is trying to convey; she can only parse the stream of sand as it comes to her. As the sand accumulates, the context becomes clearer, and perhaps the understanding does as well. Each bit of speech must be interpreted in relation to what the speaker can accomplish, namely, the relationship of the upper rim of the neck and the reinterpretation of the sound/sand delivered into what the speaker seems to have said. Without this link, the multitude of acoustic interpretations is difficult to relate to meaning, and there are many experimental results showing that the interpretation is based on articulation. The relevant articulation is not that of the listener, but that of the speaker. The hourglass's neck must constrain the possible interpretations of each bit of speech in ways that the listener can relate to a human vocal tract, and, generally, this will not be her own. If she were only able to relate sounds to her own production, speech would be useless.

Once an utterance is over (if politeness rules), it is entirely possible for the listener to flip the hourglass and become the speaker. This is parity (Liberman & Whalen, 2000). The constraints in the neck now become the basis for the speaker's confidence that the new utterance can be understood by a new listener. This is why there are so many discrepancies at higher and lower levels that are still obvious, even if we accept a crucial link between production and perception. Selection is critical for the speaker and a challenge for the listener. The flow of sand/speech can be further restricted by the speaker when the listener is struggling, through not knowing the language fully or being in poor listening conditions, or if the speaker herself is not fully fluent or, indeed, certain about what to say. (The acoustic aperture remains constant, perhaps overstretching the analogy.) However, all such adjustments require input from systems outside of the hourglass/speech. Therefore, there will always be substantial evidence for factors that ignore the link. That does not mean that the link can be eliminated, any more than an hourglass can do without a neck.

The "neck" in speech includes predisposed and learned knowledge of production effects on the acoustic, visual and tactile realization. How this is accomplished is, in my opinion, magic, but it is the magic of evolution. That there is a strong

predisposition is evident in the infant's ability to imitate (Studdert-Kennedy, 1986). Learning is clearly involved as is evident both in the maturation of the infant's speaking ability and in the differentiation of languages. That this set of predispositions and abilities is almost as complex as the other layers of language is responsible for the ongoing debate about the necessity for a link.

We no longer use hourglasses, having replaced them with more accurate and useful measurement systems. The analogy is still appealing to me, perhaps because it seems that humans will be replaced with more accurate and useful systems as well. The utility of analogies remains, and it is my hope that this one allows us to see the importance of the link between production and perception without losing sight of the many ways in which language is not restricted to that link. Communication via speech would not, I believe, exist without the link.

Figure 1: *An hourglass that (sort of) gives segments as a perceptual result. Courtesy Michael C. Stern and DALL-E.*

References

Denes, P. B., & Pinson, E. N. (1963). *The speech chain: The physics and biology of spoken language*. Garden City, NY: Anchor Press/Doubleday.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4*, 187-196.

Pardo, J. S., & Remez, R. E. (2021). On the relation between speech perception and speech production. In J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni (Eds.), *The handbook of speech perception* (2nd ed., pp. 632-655). Hoboken, NJ: John Wiley & Sons.

Studdert-Kennedy, M. (1986). Development of the speech perceptuomotor system. In B. Lindblom & R. Zetterstrom (Eds.), *Precursors of early speech: Proceedings of an International Symposium held at The Wenner-Gren Center, Stockholm, September 19–22, 1984* (pp. 205-217). New York: M. Stockton Press.

Effects of Age on Intra-syllabic Anticipatory Labialization in French.

Louise Wohmann-Bruzzo[1], Nicolas Audibert[1], Cécile Fougeron[1]

[1]Laboratoire de Phonétique et Phonologie

louise.wohmann@sorbonne-nouvelle.fr, nicolas.audibert@sorbonne-nouvelle.fr, cecile.fougeron@sorbonne-nouvelle.fr

Introduction.

Studying healthy aging is needed to understand pathological aging. The most studied effects of age concern modifications in voice's characteristics and rate. Several age-related changes have been advocated to affect speech and voice, including physiological changes, and decline in motor control and in other cognitive functions, which can affect speech planification (Sataloff & Kost 2020, Ketcham & Stelmach 2004, see Tucker *et al.* 2021 for a review). In a recent study, D'Alessandro & Fougeron (2021) found that anticipatory Vowel-to-Vowel coarticulation in French reduces with aging, and that this reduction of coarticulation could not be explained solely by a slowing down of articulation rate with age. Other possible interpretations were proposed to explain the reduction in coarticulation, including either a modification of the size of the speech plans over which anticipatory V-to-V coarticulation occurs – old speakers would not anticipate V_2 in a $CV_1.CV_2$ word if the speech plan is smaller than the size of the word –, or a reorganization of the coordination between gestures for a mode with less overlapping gestures. In the present study, we follow up on this question, by examining whether intra-syllabic coarticulation is also reduced for older speakers. To achieve this, we examine anticipatory labialization in the syllable /sy/ in French, in the productions of an older speakers' group, compared to a younger speakers' group. By the time of the conference, we plan to have analysed the productions of the 100 speakers (already recorded) and will be able to present data by chronological age, and not age-groups.

Methods.

The productions of 39 speakers, split in two age groups, have been analyzed so far: 20 younger participants between 23 and 34 y.o.a. (mean = 26.7, 10 females and 10 males), and 19 older participants, between 72 and 86 y.o.a. (mean = 78.2, 8 males, 11 females). In a story reading task, each speaker produced 3 repetitions of 7 French sentences, containing 8 monosyllabic words with an /s/ as onset, followed by a /y/ or a /i/ as nucleus. The /si/ syllables were used as baseline: they have a prevocalic [s] with no anticipation of labialization, and allow us to estimate the degree of coarticulation by observing the difference between [si] and [sy]. The [s] in a labialized context will be noted [s_y], and their unlabialized counterparts will be noted [s_i]. We obtained a total of 24 [s] (both [s_i] and [s_y]) for each speaker (936 [s] in total). In each sound file, the fricative [s] was segmented manually. Measures of the spectral Center of Gravity were taken at 10 equally spaced points of each [s]' duration, with a 10 ms Hanning window centered on the target point. An averaged value of CoG in bark, computed over the central portion of the /s/ (40% to 70% of total duration), was considered here. Mixed linear models were used to predict the CoG values according to: the following vowel ([y] or [i]), the age group, and their interaction as fixed factors, with speaker and sentence as random intercepts. Duration was also included as a continuous predictor in the model, to account for expected age-related variation in segmental duration. To investigate individual variability within each group, models by speakers were also fitted, with the following vowel as fixed factor.

Results & Discussion.

Our results show that the CoG is overall significantly higher for female than for male speakers, and this sex distinction is stronger on [s_y]. This sex difference has already been observed in the literature and could be accounted for by a combination of biological differences (palate size) and socio-phonetic factors (Fuchs & Toda 2010). Because we found an interaction of sex and the following vowel on CoG, we pursue our analysis on models made for female and male speakers separately. In both the male and the female models, as illustrated in Figure 1, there is an effect of the following vowel showing that the CoG of [s_y] is lower than that for [s_i], as predicted by an anticipation of rounding and protrusion during the /s/. Interestingly, this effect of the following vowel interacts with age: speakers in the older group do anticipate, but to a lesser extent than the speakers in the younger group. In other terms, older speakers show less difference in CoG between [s_i] and [s_y]. At an individual level, a significant lowering of the CoG in [s_y] vs. [s_i] (i.e. anticipatory labialization) is found for most speakers except 1 of the 18 male speakers, and four of the 20 female speakers. These 5 speakers who do not anticipate labialization are all over 76 years old, except for one young (23 y.o.a.) female speaker. Results by speaker are illustrated in Figure 2 for the female speakers.

These preliminary results, based on two age groups far apart in age, show that older speakers reduce anticipatory labialization within a syllable, in a similar way to what was found in D'Alessandro & Fougeron (2021) for anticipatory coarticulation across syllables. If a reduction of the size of the speech plans for older speakers was the explanation for the observed reduction in anticipation, it would mean that older speakers would be planning their speech segment by segment – which is very unlikely for typically aging adults. It seems rather that, as defended in D'Alessandro *et al.* (2020), older speakers do organize their speech in what we could call a 'clear speech mode', with little overlap between gestures.

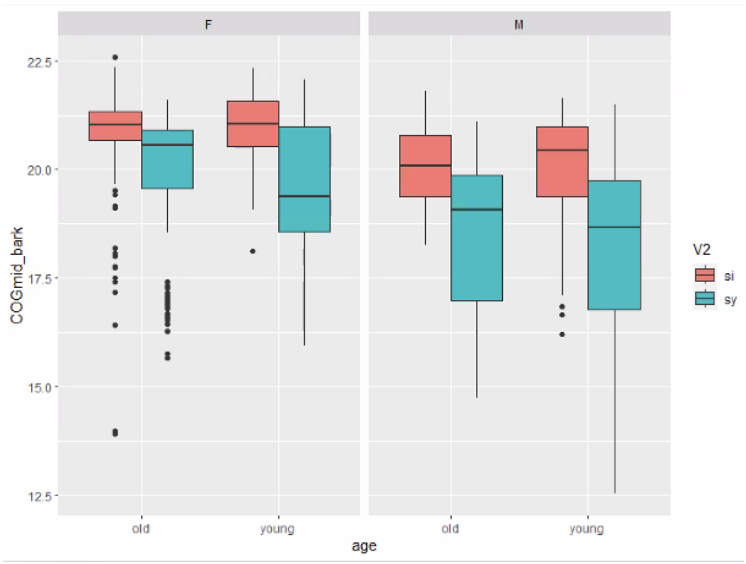


Figure 1: *CoG (in barks) according to the following vowel ([i] and [y]) and age group (older and younger) for the female speakers (left) and male speakers (right).*

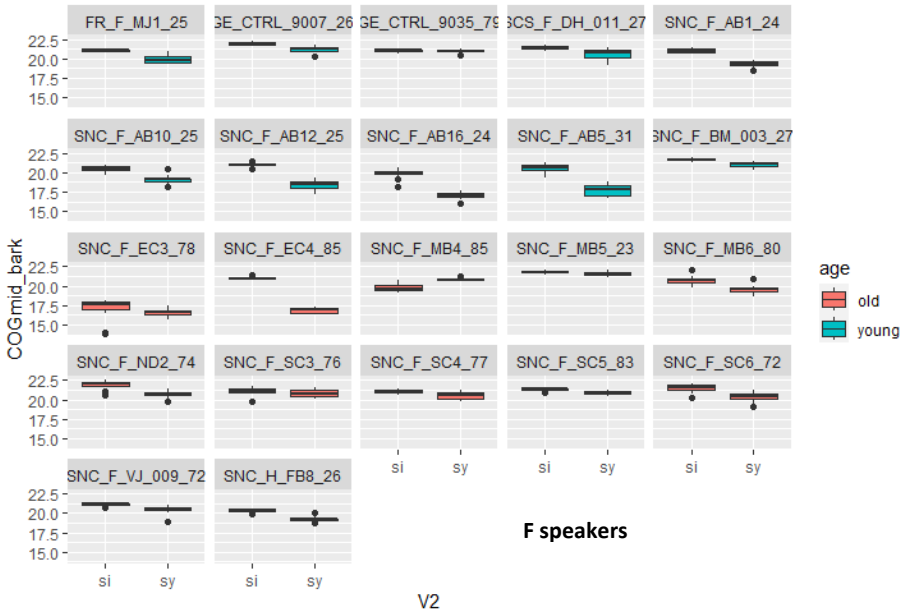


Figure 2: *differences of CoG (in barks) between [s_i] and [s_y], for each of the female speakers. Speakers with red boxplots are part of the older group, speakers with blue boxplots are part of the younger group.*

References

D'Alessandro, D., & Fougeron, C. (2021). Changes in Anticipatory VtoV Coarticulation in French during Adulthood. *Languages*, 6(181).

D'Alessandro, Daria, Angelina Bourbon, and Cécile Fougeron. (2020). Effect of age on rate and coarticulation across different speechtasks. Paper present at the 12th International Seminar on Speech Production, Virtual. December 14–18; New Haven: Haskins Press. ISBN 978-1-7360794-2-3.

Fuchs, S. & Toda, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? In S. Fuchs, M. Zygis & M. Toda (Eds.), Turbulent sounds: An interdisciplinary guide. Mouton de Gruyter. 281-302.

Hermes, A., Mertens, J. & Mücke, D. (2018). Age-related Effects on Sensorimotor Control of Speech Production. Paper present at INTERSPEECH, Hyderabad, India, September 2–6; 1526–30.

Ketcham, C. J. & Stelmach, G. E. (2004). Movement control in the older adult. *Technology for Adaptive Aging.* Washington, DC: National Academies Press US, Available online: https://www.ncbi.nlm.nih.gov/books/NBK97342/ (accessed on 15/12/2023).

Ramig, L. A. (1983). Effects of physiological aging on speaking and reading rates. *Journal of Communication Disorders*, 163. 217–26.

Sataloff, R. T., & Kost, K. M. (2020). The Effects of Age on the Voice. *Journal of Singing*, 77(1), 63-70.

Tucker, B. V., Ford, C., Hedges, S. (2021). Speech aging: Perception and Production. *WIREs Cognitive Science,* e1557.

Poster 3

03:00 – 05:00 pm

Paper	Title	Authors
64	Error profiles in the speech of children who do and do not stutter	Roaa M Alsulaiman (King Saud University)*; Zhixing Yang (University College London); Peter Howell (University College London)
161	Rhythmic predictors for speech rate variation in children who do and do not stutter	Simone Falk (University of Montreal)*; Simone Dalla Bella (University of Montreal); Mona Franke (BRAMS, University of Montreal); Ramona Schreier (Ludwig-Maximilians-University); Miriam Oschkinat (Ludwig-Maximilians-Universität München); Alicia Kluth (Ludwig-Maximilians-University); Georg Thum (Ludwig-Maximilians-University); Ingeborg Mayer (speech therapist); Philip A Hoole (Institute of Phonetics, Munich University)
78	Investigation of Vibrational Frequency of Canine Vocal Folds Using a Two-Way Fluid-Solid Interaction Analysis	Abolfazl Mohammadi Gorjaei (university of Tehran)*; Mohammad Ali MAN Nazari (University of Tehran); Asghar Afshari (university of Tehran); Saeed Farzad-Mohajeri (university of Tehran); Pascal Perrier (Grenoble INP)
88	The effect of concurrent linguistic and nonlinguistic task on speech motor performance in Parkinson's Disease	Hanna Rakhangi (University of Toledo); Dema Herzallah (University of Toledo); Olumide Oyebode (University of Toledo); Jennifer Peterson (University of Toledo); Caroline Menezes (University of Toledo)*
182	Progressive speech type adaptation to distance: frequencies of /a/ as a case study	Julien meyer (Université Grenoble Alpes, CNRS, GIPSA-Lab)*; Adèle Denis (Université de Rennes); Laure Dentel (The World Whistles Research Association)
136	Predicting articulatory landmarks with critically-damped oscillators and General Tau Theory	Christopher A Geissler (Carleton College)*; Jyothiraditya Nellakra (Carleton College)
124	Exploring Vowel Reduction in French Casual Conversations	Kübra Bodur (LPL, Aix-Marseille Université)*; Corinne Fredouille (Avignon Université - LIA); Christine Meunier (Aix-Marseille Université - LPL)
149	Speaking style influence on vowel length opposition in Jordanian Arabic	Mohammad Abuoudeh (Al Hussein Bin Talal University)*; Jalal Al-Tamimi (Université Paris Cité); Olivier Crouzet (Université de Nantes)
13	Spatiotemporal coordination of tongue dorsum characterizes the voicing contrast of American English bilabial coda obstruents	Daejin Kim (University of New Mexico)*
231	Probing the impact of musical training on the temporal malleability of timing in French	Nicole Benker (Ludwig-Maximilians-Universität München); Miriam Oschkinat (Ludwig-Maximilians-Universität München)*; Philip A Hoole (Institute of Phonetics, Munich University); Simone Falk (University of Montreal); Simone Dalla Bella (University of Montreal)
92	Acoustic Analysis of Fricatives in Lushootseed	Ted Kye (University of Washington)*
112	Dentition and Articulations of Mandarin Rhotic /ɻ/ and Retroflex /ʂ/: A Preliminary Study	YUNG-HSIANG SHAWN CHANG (National Taipei University of Technology)*; Jeyang Jau (O2Win Dental Clinic)
176	Prosodic variation of kibushi dialects (Mayotte, France) via the reading of The Little Prince : a pilot study	Ahamada KASSIME (Paul Valéry University Montpellier 3)*
121	EEG dynamics and source identification during air volume reduction induced by a long utterance	Said-Iraj Hashemi (LPP CNRS , LNMB ULB)*; Guy Cheron (LNMB); Didier Demolin (LPP CNRS); Ana Maria Cebolla Alvarez (LNMB ULB)
129	Laterals in simplex vs. complex syllable codas: a comparison of four languages	Anisia Popescu (Université Paris Saclay - LISN)*; Ioana Chitoran (Université Paris Cité)
217	Some Effects of Framerate on Gesture Detection in Tongue Ultrasound	Pertti Palo (Indiana University)*; Steven M. Lulich (Indiana University)
102	Why do palatographic data have to be taken seriously?	Yury Makarov (Institute of Linguistics, RAS; University of Cambridge)*
111	Laryngeal changes associated with Guttural consonants in Levantine Arabic	Jalal Al-Tamimi (Université Paris Cité)*
82	Vocal Motor Control During Exposure to Oscillating Pitch Changes	Rita Bishai (Wilfrid Laurier University)*; Jeffery Jones (Wilfrid Laurier University); Nichole Scheerer (Wilfrid Laurier University)
138	Past sensory-prediction-error and motor-compensation biases asymmetrically mediate speech motor control	Yuhan Lu (NYU Shanghai)*; Tingting Wang (Minerva University); Xing Tian (New York University Shanghai)
106	The Acoustics of Vowel Sequences in Five Romance Languages	Johanna Cronenberg (Université Paris Cité)*; Lori Lamel (LISN); Ioana Chitoran (Université Paris Cité)
194	Does Cross-Word Resyllabification Modify Word Cohesion in French?	Alice Yildiz (Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle))*; Anne Hermes (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS & Sorbonne Nouvelle, Paris); Cecile FOUGERON (LPP)
8	Speech onset kinematics predict sentence level variability in adults who stutter	Daniel Aalto (University of Alberta)*; Torrey Loucks (Jacksonville University)
66	Acoustic characteristics of narrative elements in infant-directed speech	Anna Kohari (HUN-REN Hungarian Research Centre for Linguistics)*; Katalin Mády (HUN-REN Hungarian Research Centre for Linguistics); Uwe D. Reichel (HUN-REN Hungarian Research Centre for Linguistics); Veronika Harmati-Pap (HUN-REN Hungarian Research Centre for Linguistics); Bence Kas (Eötvös Loránd University, MTA-ELTE Language-Learning Disorders Research Group, HUN-REN Hungarian Research Centre for Linguistics); Sarolta Murányi (ELTE Department of Applied Linguistics and Phonetics)
227 (Remote)	Lingual and epilaryngeal articulation of vowels in Mundabli	Matthew Faytak (University at Buffalo)*; Mariana Quintana Godoy (University at Buffalo); Tianle Yang (University at Buffalo)

225 (Remote)	Decoding orofacial signals beyond sight: A study of expressive faces and whispered voices in German	Nasim Mahdinazhad Sardhaei (Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS))*; Marzena Zygis (Leibniz-ZAS); Hamid Sharifzadeh (Unitec Institute of Technology)
140	Linking levels of prosodic structure to modulatory activity in speech production	Leonardo Lancia (Laboratoire Parole et Langage (CNRS/Aix-Marseille Université))*; Jiayu Li (Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle)); Caterina Petrone (LPL); Louis Goldstein (University of Southern California)
72	Sibilant contrast production by bilingual speakers of Quanzhou Southern Min and Mandarin	Caihong Wang (Université Paris Cité)*; Ioana Chitoran (Université Paris Cité); Alexander Martin (University of Groningen)
39	Encoding of speech modes with varying articulatory and phonatory properties; an ERP investigation	Bryan Sanders (University of Geneva)*; Marina Laganaro (University of Geneva)
29	Velum lowering and tongue tip constriction in German VN sequences across different speech styles	Esther Kunay (Institute of Phonetics, Munich University)*; Lillian van Bressensdorf (Institute of Phonetics, Munich University); Philip A Hoole (Institute of Phonetics, Munich University); Jonathan Harrington (Institute of Phonetics, Munich University); Dirk Voit (Max Planck Institute for Multidisciplinary Sciences); Jens Frahm (Max Planck Institute for Multidisciplinary Sciences)
34	Liquids in Upper Sorbian: an MRI examination	Juliusz Cęcelewski (Laboratoire de Phonétique et Phonologie), Phil J Howson (Leibniz-Zentrum Allgemeine Sprachwissenschaft)*; Peter Birkholz (TU Dresden); Cédric Gendrot (LPP)
71	Temporal processing of sentence production in Parkinson's disease	Fatemeh Mollaei (University of Reading)*
238	Subthalamic and cortical encoding of syllable sequences	Andrew Meier (Boston University)*; Alan Bush (Massachusetts General Hospital); Mark Richardson (Massachusetts General Hospital); Frank H. Guenther (Boston University)
141	Allophones of Korean /l/: a classification using EMA	Kye Shibata (National Tsing Hua University)*; Feng-fan Hsieh (National Tsing Hua University); yueh-chin chang (NTNU)
50	A framework for modeling the rhythmic organisation of speech and the impact of perceptual cues on production	Méllen Guillaume (Université de Lille)*; Julien Diard (CNRS - Laboratoire de Psychologie et NeuroCognition), Anahita Basirat (University of Lille)
89	Effect of therapeutic boxing on actual vs individual awareness of speech production in Parkinson's disease	Geraldine Menezes (University of Toledo)*; Beth Ann Hollenich (University of Toledo)
41	Orofacial somatosensory abilities in speech motor control: a conceptual framework	Jean-Francois Patri (Université catholique de Louvain)*; Gilles Vannuscorps (Université catholique de Louvain)
67	Mapping Speech and Facial Muscles: Using Generalized Additive Modeling to Understand Speech Production through Electromyography	Inge Salomons (University of the Basque Country)*; Inma Hernáez (University of the Basque Country), Eva Navas (University of the Basque Country (UPV/EHU)), Martijn Wieling (University of Groningen)
226	Intensity downtrends in Embosi intonation	Tubin Zhang (USC)*; Annie Rialland (Laboratoire de Phonétique et Phonologie); Yijing Lu (University of Southern California); Sarah Harper (UCSF); Louis Goldstein (USC)
37	Fundamental frequency as an acoustic cue to phonological phrase boundary in Spanish	Mario Casado-Manrique (UNED)*
235	Comparison of Velum Control in /an/-rime Words between Chengdu Dialect and Standard Mandarin	Sishi Liao (Institute for Phonetics and Speech Processing (IPS), LMU Munich)*; Philip A Hoole (Institute of Phonetics, Munich University); Jonathan Harrington (Institute for Phonetics and Speech Processing (IPS), LMU Munich)
142	Speech production variability in children learning to read	Sandy Abu El Adas (Basque Center On Cognition, Brain and Language)*; Marie Lallier (Basque Center On Cognition, Brain and Language)
85	Effects of an ultrasound biofeedback session on maximal tongue movements	Eija Aalto (Ecole de technologie supérieure)*; Minaru Yoshida (Ecole de technologie supérieure), Lucie Menard (Université du Québec à Montréal); Waldir Cardoso (Concordia University); Catherine Laporte (Ecole de technologie supérieure)
119	The role of face and head movement in the production of lexical tones in Cantonese	João Vítor Passamai de Menezes (TU Dresden)*; Hani C Yehia (UFMG - Universidade Federal de Minas Gerais); Maria Mendes Cantoni (Federal University of Minas Gerais); Adriano Vilela Barbosa (UFMG); Denis Burnham (Western Sydney University)
69	Schwa optionality in verbal inflection in German: the effects of stress and phonetic context	Marie-Theres Weidgenter (Humboldt-Universität zu Berlin)*
118	Self-supervised learning of the relationships between speech units, gestures and sounds using vocal imitation	Marc-Antoine Georges (GIPSA-lab, Grenoble Alpes Univ); Marvin Lavechin (GIPSA-lab, Grenoble Alpes Univ); Jean-Luc Schwartz (GIPSA-lab, Grenoble Alpes Univ); Thomas Hueber (CNRS/Grenoble Alpes Univ)*
151	Physiological constraints underlying the variation of labial stop intensity and spectrum	Maëva GARNIER (GIPSA-lab)*; Thibault Cattelain (GIPSA-lab); Christophe Savariaux (GIPSA-lab), Pascal Perrier (Gipsa-lab, Grenoble INP, Université Grenoble Alpes)
155	Are glottalic mechanisms in Human Beatboxing really glottalic ?	Alexis Dehais-Underdown (LPP CNRS)*; Lise BUCHMAN (LPP); Didier Demolin (LPP CNRS); Pierre André Vuissoz (IADI & INSERM); Marc Fauvel (CIC-IT & INSERM); Yves Laprie (CNRS/Loria); Jacques Felblinger (IADI & CIC-IT & INSERM)

Error profiles in the speech of children who do and do not stutter

Roaa Alsulaiman[1,2], Zhixing Yang[2], Peter Howell[2]
King Saud University, University College London

The production of spontaneous speech is not always without errors. In fact, speech production is an extremely complex process that depends on the precise coordination of laryngeal, orofacial, and respiratory muscles (Simonyan et al., 2016). The flow of speech may be disrupted if the speaker detects an error in any of these articulators. These disruptions are manifest as speech disorders such as stuttering. The incidence of stuttering given by Andrew and Harris (1964) was 4.9%, and it has been emphasized that early identification and interventions of any potential speech dysfluencies facilitates recovery (Howell, 2011).

The type of disruption in stuttering is different from the normal disfluencies experienced by fluent speakers. There are three main types of stuttered disfluencies; sound and syllable repetition, prolongation of voiced or voiceless sounds and failure within an attempt to produce a sound, known as "breaks". These symptoms are used to assess stuttering in the stuttering severity instrument (SSI-3; Riley, 1994), which is the most widely used stuttering screening instrument. A more convenient method of assessing stuttering than the SSI-3 is the non-word repetition (NWR) task. Literature has shown that children who stutter (CWS) may have NWR deficits due to phonological processing impairments since CWS are less accurate on NWR tasks compared to children who do not stutter (CWNS) (Anderson & Wagovich, 2010). Other source of fluency issues affect NWR performance. For instance, Windsor et al (2010) found that typical children performed better in NWR tests made in their first language compared to one designed for another language. To avoid language biases, Howell et al (2017) designed the universal NWR (UNWR), which allows equitable testing for 20 different languages. UNWR could be used to identify CWS irrespective of language spoken.

Howell et al (2017) looked at whether UNWR can identify CWS and CWNS, however, there is limited research about the influence of linguistic factors on the occurrence of stuttering. Such information may provide better understanding about the mechanisms involved in the production of stuttered speech. Brown (1945) found that stuttering is often influenced by linguistic factors including initial phoneme and word length and. Moreover, multisyllables is an aspect of phonological difficulty that may influence stuttering (Throneburg et al, 1994). Literature has also found that the majority of stuttering events occur on the onset of words, with Natke et al (2003) finding that 97.8% of stuttering events occur on the first syllable of words.

Thus, the aim of this study was to investigate whether the phonological difficulty (measured by word length of non-words impacts UNWR performance in CWS and CWNS. Additionally, we investigated whether CWS show more stuttering on word onsets. Finally, in-depth exploratory analyses of consonant substitution, deletion, and insertion error profiles are conducted.

Participants, materials and study design

26 participants were included (13 CWS, Mage=11.77; SD=4.73, and 13 typically developing CWNS, Mage=11.62; SD=4.19). Of the 13 CWS, 12 had a confirmed diagnosis by a speech and language therapist. The CWS group comprised 10 males and 3 females; and the CWNS comprised 9 females. All participants reported English as their dominant language, including 17 monolingual English speakers and 9 multilingual speakers.

Stimuli included the 2 syllable and 3 syllable nonwords used in the UNWR task (Howell et al., 2017). The non-words were pre-recorded and spoken by a professional phonetician to ensure that there is no variation in delivery. The presence or absence of stuttering was treated as the between-subjects variable. The within-subjects variable was the manipulation of nonword length (i.e. The increase in syllables number), which was tested at two levels with syllable lengths ranging between 2-3 syllables.

Nonwords were orthographically transcribed and compared phoneme-by-phoneme to each target nonword. Accuracy proportions were determined based on the total number of repeated nonwords. Then, consonant errors were categorized into three types: substitutions, deletions or insertions (*Table 1* has examples).

Table 1
Examples of errors on the UNWR task

Error Type	Target Response	Error Response
Substitution	**f**lon-tren-drut	**p**lon-tren-drut
Deletion	flon-tren-drut	flon-t()en-drut
Insertion	flon-tren-drut	flon-tren-dru**n**t

Results and discussion

A mixed-factorial analysis of covariance (ANCOVA) was conducted to assess whether CWS performed differently depending on the nonwords' phonological difficulty, and whether their overall accuracy (i.e. proportion of accurately repeated nonwords) on the UNWR task differed from CWNS. Overall, participants performed worse when repeating 3-syllable nonwords (M = .47) compared to 2-syllable nonwords (M = .56), but not to a statistically significant extent, F(1,22) = 1.23, p = .279, ηp2 = 5.3%. Significant differences were found between CWS and CWNS across the combined nonword lengths, F(1,22) = 22.64, p < .001, ηp2 = 50.7%. However, There was no significant interaction between phonological difficulty and group, such that both CWS and CWNS performed proportionally worse at the increased syllable length, F(1,22) = .369, p = .550, ηp2 = 1.6%.,

A mixed-factorial analysis of variance (ANOVA) was conducted to see whether CWS exhibits more stuttering in the first syllables compared to final syllables of three-syllable. It was found that CWS, were more likely to stutter when uttering the first syllable of a three-syllable nonword. In this case, a significant main effect of order of utterance was observed, $F(1,24) = 5.27$, $p = .031$. The same analysis was conducted on nonwords of two syllables, and it was found that CWS were more likely to stutter when uttering the first syllable of a two-syllable nonword, but no significant main effect of order of utterance was observed, $F(1,24) = 1.27$, $p = .27$. These findings partially align with Howell and Au-Yeung's (2002) EXPLAN theory of serial ordering, which proposes that disfluencies are more likely to occur at the word-onset, especially for more complex phonological sequences.

With respect to error types, the results showed that CWS were significantly more likely to produce substitution errors than their fluent peers, across both levels of phonological difficulty. This indicates that CWS may have difficulty holding detailed phonological representations in working memory, leading to inaccurate phoneme retrieval during the UNWR task. CWS may have difficulty breaking down speech sounds into individual phonemes and instead process them as larger and less detailed units. This increases the likelihood for target phonemes to be replaced by phonologically adjacent phonemes. However, this can also lead to difficulty in producing fluent spontaneous speech, as CWS struggle with the precise motor-programming of individual speech sounds (Wolk et al., 1993; Howell, 2004).

Implications and future work

The current study provides compelling evidence supporting the use of the UNWR task as an effective screening tool for stuttering. Furthermore, the exploratory analyses shed light on various aspects of phonological processing, suggesting that error profiles could be further examined in future research to gain a better understanding of stuttering. Data collection is ongoing and results from a larger number of children will be reported at the conference.

References

Anderson, J. D., & Wagovich, S. A. (2010). Relationships among linguistic processing speed, phonological working memory, and attention in children who stutter. *Journal of Fluency Disorders*, *35*(3), 216–234. https://doi.org/10.1016/j.jfludis.2010.04.003

Andrews, G., & Harris, M. (1964). The syndrome of stuttering. Spastics Society Medical Education.

Brown, S. F. (1945). The loci of Stutterings in the speech sequence. *Journal of Speech Disorders*, *10*(3), 181–192. https://doi.org/10.1044/jshd.1003.181

Howell, P., Soukup-Ascencao, T., Davis, S., & Rusbridge, S. (2011). Comparison of alternative methods for obtaining severity scores of the speech of people who stutter. Clinical Linguistics & Phonetics, 25(5), 368-378

Howell, P., Tang, K., Tuomainen, O., Chan, S. K., Beltran, K., Mirawdeli, A., & Harris, J. (2017). Identification of fluency and word-finding difficulty in samples of children with diverse language backgrounds. *International Journal of Language & Communication Disorders*, *52*(5), 595–611. https://doi.org/10.1111/1460-6984.12305

Riley, G. (1994). *The Stuttering Severity Instrument for Adults and Children (SSI-3)* (3rd ed.). Austin, TX: PRO-ED.

Simonyan, K., Ackermann, H., Chang, E. F., & Greenlee, J. D. (2016). New developments in understanding the complexity of human speech production. Journal of Neuroscience, 36(45), 11440-11448.

Throneburg, R. N., Yairi, E., & Paden, E. P. (1994). Relation between phonologic difficulty and the occurrence of disfluencies in the early stage of stuttering. Journal of Speech, Language, and Hearing Research, 37(3), 504–509. https://doi.org/10.1044/jshr.3703.504

Windsor, J., Kohnert, K., Lobitz, K. F., & Pham, G. T. (2010). Cross-language nonword repetition by bilingual and monolingual children. *American Journal of Speech-Language Pathology*, *19*(4), 298–310. https://doi.org/10.1044/1058-0360(2010/09-0064)

Yairi, E., & Ambrose, N. (2013). Epidemiology of stuttering: 21st century advances. *Journal of Fluency Disorders*, *38*(2), 66–87. https://doi.org/10.1016/j.jfludis.2012.11.002

Rhythmic predictors for speech rate variation in children who do and do not stutter

Simone Falk[1], Simone Dalla Bella[1], Mona Franke[1,2], Ramona Schreier[2], Miriam Oschkinat[2], Alicia Kluth[2] , Georg Thum[2], Ingeborg Mayer, Philip Hoole[2]

[1]BRAMS laboratory, University de Montréal, Montréal, Canada
[2]Institute for Phonetics and Speech processing, Ludwig-Maximilians-University, Munich, Germany
Simone.falk@umontreal.ca

Introduction. Altered speech and articulation rate are prominent markers of developmental speech motor disorders in children and adolescents. In stuttering, for instance, both speech and articulation rate tend to be slower in children because of actual dysfluencies as well as altered speech motor control (e.g., Hall et al., 1999). Techniques for modulating speech rate are widely used in stuttering therapy in order to achieve better stuttering management and to enhance fluency (e.g., the Camperdown program, O'Brian et al., 2003). Hence, beyond maturation with age, variation in speech rate in children who stutter can have different individual sources, such as stuttering severity, stuttering management and general speech motor processes. In the present contribution, we pursue the question whether variation in speech and articulation rates in children who stutter can also be predicted by musical rhythmic abilities. Some have argued that training rhythmic abilities could benefit altered speech and language processes in neurodevelopmental disorders such as stuttering (e.g., Fujii & Wan, 2014). Performing musical rhythms is a way to train sensorimotor-coupling (Fiveash et al. 2021), a capacity that is needed in music and speech to precisely time complex motor patterns to generate fast-paced auditory sequences. In stuttering, sensorimotor networks in the brain show alterations which lead to less efficient auditory-motor integration and ultimately, stuttering symptoms (e.g., Chang & Guenther, 2020). Therefore, rhythmic synchronization tasks relying on precise auditory-motor skills could be a good way to test the importance of auditory-motor skills for speech motor development in stuttering. For example, we have previously found that children and adolescents with more severe stuttering show less precision and consistency in rhythmic auditory-motor synchronization tasks (e.g., tapping to music; Falk et al., 2015). In sum, we hypothesize that children who stutter with better auditory-motor synchronization skills could also be those showing less impact of stuttering on their speech rate.

Methods. To explore this hypothesis, we tested 98 children and adolescents between 9-20 years (mean age=13, 6 female participants), of whom 49 do and 49 do not stutter, on a battery of auditory-motor synchronization tasks (e.g., Dalla Bella et al, 2017). In one set of the tasks, participants were asked to tap their finger to non-verbal as well as to verbal rhythmic auditory stimuli, such as a metronome, music, syllables, lists of words and highly rhythmical sentences. In another set of tasks, participants synchronized syllables and words to a metronome beat. We assessed speech and articulation rates of spontaneous speech and of reading. Spontaneous rates were evaluated in a semi-structured interview with the experimenter. To assess reading rates, participants read an excerpt of a popular children's book. Generally, rates were estimated in syllables per second for an excerpt of ~1 minute of interview/read text. Dysfluencies (e.g., reading errors, hesitations, stuttering) were identified in both excerpts. To calculate articulation rates, pauses (> 250 ms) and dysfluencies were subtracted from the overall sentence duration. Linear models were fit to the data, with sets of predictors derived from synchronization performance (consistency and timing of taps to syllables/metronome; taps with music/speech; taps with words; speaking to a metronome), as well as age and group. Finally, stuttering severity (SSI-3, Riley, 2003) was assessed by trained clinicians.

Results. Our results showed a group by age interaction as a predictor for speech and articulation rates in both spontaneous speech and reading. Only the control group showed faster rates with age, while participants who stutter did not, speaking generally slower than the control group. Participants who stutter showed a strong relation between stuttering severity (SSI) and speech and articulation rates. Beyond age and group, articulation rates in spontaneous and read speech were also predicted by how consistent participants tapped with complex stimuli such as music and speech, independently of stuttering. Importantly, speaking with a metronome predicted articulation rates in spontaneous and read speech differently in children and adolescents who do and do not stutter (Fig. 1). The more children who stutter were consistent when synchronizing vowels of syllables and words to a metronome, the more their articulation rate approached the control group's rates. The result also holds when stuttering severity is taken into account. Similar results were found for spontaneous speech rate, but not for read speech rate. No relation between consistency of metronome speech and rates was found in the control group. Interestingly, higher consistency when speaking with a metronome also predicted lesser dysfluencies in spontaneous and read speech in participants who stutter. Finally, we observed less dysfluencies in participants who stutter who timed their finger taps closer to the beat or vowel in metronome / syllable tapping.

Discussion. These results show that performance in auditory-motor synchronization tasks can indeed predict articulation rates and, to a certain extent, speech rates, in spontaneous speech and reading of children and adolescents. In line with our hypothesis and recent frameworks on the beneficial effects of rhythmic capacities in neurodevelopmental speech disorders, participants with stuttering showed faster rates and less dysfluencies when being better synchronizers in certain rhythmic tasks. Speaking with a metronome as well as tapping to a metronome / syllables were those tasks providing the most powerful predictors for participants who stutter. Tapping with complex stimuli such as music and a metronome was predictive for all participants independently of stuttering. Overall, these results suggest that speech motor development is related to more general sensorimotor timing or rhythm skills. In particular, we were able to show the relevance for stuttering, a speech motor disorder originating from altered communication between auditory and motor networks in the brain. Future research should continue to investigate potential causal relations, via auditory-motor and the timing network in the brain, and differences between developmental populations with and without alterations in these neural ressources.

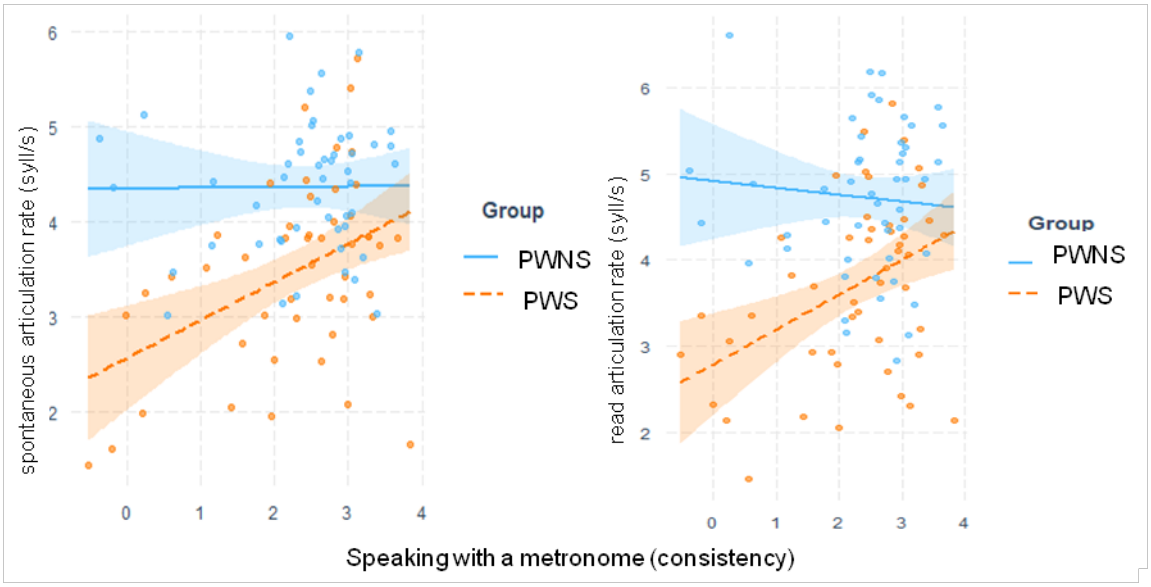


Figure 1. *Slopes of articulation rates in read and spontaneous speech (y-axis) predicted by consistency (x-axis) when speaking with a metronome (i.e., the variability in timing vowels of syllables and words to the beat) in participants who do (PWS) an do not stutter (PWNS, blue). Higher consistency predicts faster rates only in PWS.*

References

Chang, S-E., & Guenther, F. H. (2020). Involvement of the cortico-basal ganglia-thalamocortical loop in developmental stuttering. *Frontiers in Psychology*, *10*, Article 3088.

Dalla Bella, S., Farrugia, N., Benoit, C. E., Begel, V., Verga, L., Harding, E., & Kotz, S. A. (2017). BAASTA: Battery for the Assessment of Auditory Sensorimotor and Timing Abilities. *Behavior research methods*, *49*(3), 1128–1145.

Falk, S., Müller, T., & Dalla Bella, S. (2015). Non-verbal sensorimotor timing deficits in children and adolescents who stutter. *Frontiers in Psychology*, *6*, Article 847.

Fiveash, A., Bedoin, N., Gordon, R. L., & Tillmann, B. (2021). Processing rhythm in speech and music: Shared mechanisms and implications for developmental speech and language disorders. *Neuropsychology*, *35*(8), 771–791.

Fujii, S., & Wan, C. Y. (2014). The role of rhythm in speech and language rehabilitation: The SEP hypothesis. *Frontiers in Human Neuroscience*, *8*, Article 777.

Hall, K. D., Amir, O., & Yairi, E. (1999). A Longitudinal Investigation of Speaking Rate in Preschool Children Who Stutter. *Journal of Speech, Language, and Hearing Research, 42*, 1367-1377. doi:10.1044/jslhr.4206.1367.

O'Brian, S., Onslow, M., Cream, A., & Packman, A. (2003). The Camperdown programm: Outcomes of a new prolonged-speech treatment model. *Journal of Speech Language and Hearing Research, 46*(4), 933-946.

Investigation of Vibrational Frequency of Canine Vocal Folds Using a Two-Way Fluid-Solid Interaction Analysis

Abolfazl Mohammadi Gorjaei[1], Mohammad Ali Nazari[1], Asghar Afshari[1], Saeed Farzad-Mohajeri[2], Pascal Perrier[3]

[1] School of Mechanical Engineering, University of Tehran, Tehran, Iran

[2] Department of Surgery and Radiology, Faculty of Veterinary Medicine, University of Tehran, Tehran, Iran

[3] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

amohammadig77@ut.ac.ir, manazari@ut.ac.ir, afsharia@ut.ac.ir, saeedfarzad@ut.ac.ir, Pascal.Perrier@grenoble-inp.fr

Introduction. Speech is an integral component of human communication, requiring the coordinated efforts of various organs to produce sound (Titze & Alipour, 2006). The glottis region, a key player in voice production, assumes a crucial role in this intricate process. As air, emanating from the lungs in a confined space, interacts with the vocal folds (VFs) within the human body, it gives rise to the creation of voice (Alipour & Vigmostad, 2012). Understanding the mechanical intricacies of this process is very important. Studying VFs in vivo situations is hard work. However, the orientation, shape and size of VFs fibers have been extracted with synchrotron X-ray microtomography. (Bailly *et al*., 2018)

The investigation of mechanical properties of both human and animal VFs has been carried out through various methodologies in the literature. The mechanical properties of VFs have been studied using the uniaxial extension test (Alipour & Vigmostad, 2012) assuming a linear behavior, while the nonlinearity and anisotropy of VFs has been determined using a multiscale method as in Miri *et al*. (2013). Pipette aspiration has also been used to extract in vivo elastic properties of VFs (Scheible *et al*., 2023). Mechanical behavior of VFs layers in tension, compression and shear has been studied. (Cochereau *et al*., 2020). Fluid-structure interaction (FSI) simulations provide a valuable tool to gain a deeper understanding of voice production (Ghorbani *et al*. 2022). These simulations allow us to model the dynamic interplay between the VFs and air. Our research focuses on investigating the mechanical properties of canine vocal folds and utilizing these findings in an FSI simulation. Through this simulation, we aim to unravel how these mechanical properties affect voice production.

Methods. To investigate the mechanical properties of canine VFs, an in vitro study was conducted involving 6 mixed-breed dogs. The samples were harvested from canine cadavers euthanized for reasons unrelated to this study. In the following, the VFs were harvested and tested upon 3-4 hours post-animal sacrifice.

Experimental trials were carried out using the STM-1 device (SANTAM Co.), equipped with a 100 kg load cell. Seven uniaxial tensile tests were done on each sample, with displacement rates of 1, 5, 10, 20, 40, 60, and 120 mm/min. The very slow rate of 1 mm/min was chosen to assess only elastic properties eliminating viscosity effects. Various hyperelastic models were used to fit the experimental data. Subsequently, for each model, both the mean and standard deviation (SD) were determined for the hyperelastic model parameters and their residuals.

For FSI analysis we used a simplified laryngeal model as a hollow cylinder with a diameter of 50 mm and a thickness of 3 mm. The overall length of the larynx was set at 100 mm. The VFs were modeled as a circular disc with a small elliptical fissure in the midst of the cylinder section. Boundary conditions were established based on pressure differentials, with the inlet gauge pressure set at 1200 Pa and the relative pressure at the outlet set to 0. To account for the turbulent nature of airflow within the larynx, we employed the K-epsilon method to solve the motion differential equations in a two-way fluid-structure interaction simulation using ANSYS FLUENT 2021.

This approach enabled us to investigate how the acquired mechanical properties of canine vocal folds affect the FSI simulations during phonation, resulting in a more comprehensive understanding of their impact. To determine the vibrational frequency of VFs, we calculated the time it took to reach maximum displacement and then quadrupled this value to obtain the period of vibration.

Results. The Yeoh 2nd order model emerged as the most fitting choice with its strain energy density function:

$$\psi = C_{10}(I_1 - 3) + C_{20}(I_1 - 3)^2$$

where C_{10} and C_{20} represent the model parameters and I_1 denotes the first invariant of the Cauchy-Green strain tensor. The associated material constants derived from this model and averaged between all six canine samples were found as: C_{10}=195.8±139.7 kPa, C_{20}=6765.2±1469.4 kPa.

Stevens (2000) has expressed the deformation of VFs in 8 steps. The vibration of the simplified VFs in our simulations shows the first four steps and thus indicates the half cycle of the VFs vibration. Therefore, the time duration of oscillating behavior of VFs in Y direction was used to compute the vibrational frequency of the simplified VFs. The vibrational

frequency extracted from these results show a frequency about 125 Hz. This frequency has been calculated by inversing the twice the time taken for maximum displacement. Our findings indicate that the simplified larynx model, incorporating the extracted mechanical properties, displays a compelling behavior of voice production.

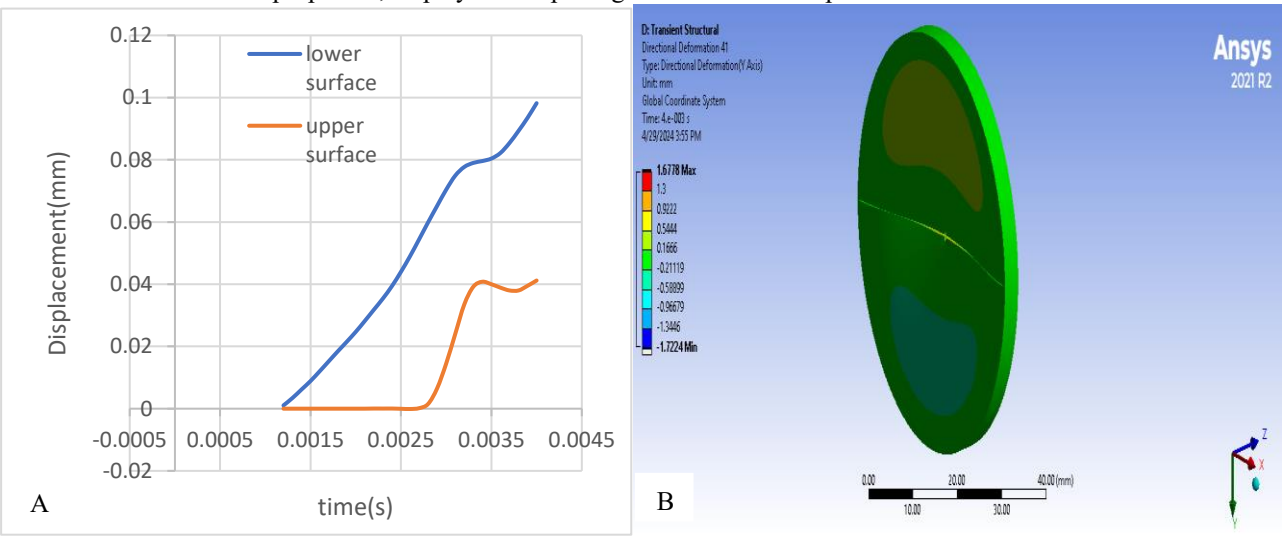


Figure 1 : *The results of the FSI simulation A- Relation of the directional displacements of the VFs in Y axis over time for two points B- the deformed shape of VFs in time=0.004s*

Figure 1-A, (the left panel) shows the displacements in Y axis direction over time for two points in z direction on the surface of fissure located on upstream and downstream parts of model. As it can be seen the downstream point (blue curve) is opened while the upstream point (red curve) is still closed. In the following the upstream point is opened and the air flows through the opened fissure. **Figure 1-B** indicates the maximum deformation of the model when the contact between fissures opens. The frictionless contact has been defined in the surface of fissure (XZ plane).

Discussion. The mechanical properties of the VFs in the previous study of our group on speech production via FSI simulation were based on the cadaver studies by Alipour, F., & Vigmostad, S. (2012). A hyperelastic model (first order Ogden) was used to fit their experimental results (Ghorbani *et al.,* 2022). In the current research, the fresh canine VFs were used to have the VFs mechanical properties. The number of specimens and the control on strain rate in our tensile experiments gives promising behavior than the former one. At the same time, we are aware that the viscoelastic properties of VFs play an essential role in their behavior. The extracted experimental data of samples let us studying this behavior which is the subject of ongoing research.

Our results for the frequency of canine VFs are close to the results of Solomon et al. (1995) which reported the frequency of growl to be equal to 112 Hz. These results indicate that our simplified simulation and the derived mechanical properties exhibit promising accuracy in predicting vocal fold behavior.

References

Alipour, F., & Vigmostad, S. (2012). Measurement of vocal folds elastic properties for continuum modeling. *Journal of voice: official journal of the Voice Foundation*, 26(6), 816.e21–816.e29.

Bailly, L., Cochereau, T., Orgéas, L., Henrich Bernardoni, N., Rolland du Roscoat, S., McLeer-Florin, A., Robert, Y., Laval, X., Laurencin, T., Chaffanjon, P. and Fayard, B. (2018). 3D multiscale imaging of human vocal folds using synchrotron X-ray microtomography in phase retrieval mode. *Scientific reports*, *8*(1), 14003.

Cochereau, T., Bailly, L., Orgéas, L., Bernardoni, N. H., Robert, Y., & Terrien, M. (2020). Mechanics of human vocal folds layers during finite strains in tension, compression and shear. *Journal of biomechanics*, *110*, 109956.

Ghorbani, O., Afshari, A., Perrier, P., Nazari, M.A. (2022). Fluid-Structure Interaction analysis of human's vocal folds and the internal airflow using an asymmetric glottis opening. 9th World Congress of Biomechanics, 2022 Taipei. Miri, A. K., Heris, H. K., Tripathy, U., Wiseman, P. W., & Mongeau, L. (2013). Microstructural characterization of vocal folds toward a strain-energy model of collagen remodeling. *Acta biomaterialia*, 9(8), 7957-7967.

Scheible, F., Lamprecht, R., Schaan, C., Veltrup, R., Henningson, J. O., Semmler, M., & Sutor, A. (2023). Behind the complex interplay of phonation: Investigating elasticity of vocal folds with pipette aspiration technique during ex vivo phonation experiments. *Journal of Voice* (In press).

Solomon, N. P., Luschei, E. S., & Liu, K. (1995). Fundamental frequency and tracheal pressure during three types of vocalizations elicited from anesthetized dogs. *Journal of Voice*, *9*(4), 403-412.

Stevens, K. N. (2000). Source mechanisms. In Acoustic Phonetics (pp. 55-127). The MIT Press.

Titze, I. R., & Alipour, F. (2006). The myoelastic aerodynamic theory of phonation. Iowa City, IA: National Center for Voice and Speech.

The effect of concurrent linguistic and nonlinguistic task on speech motor performance in Parkinson's Disease

Hanna S. Rakhangi, Dema M. Herzallah, Olumide E. Oyebode, Jennifer Peterson and Caroline Menezes

University of Toledo

hanna.rakhangi@rockets.utoledo.edu, olumide.oyebode@rockets.utoledo.edu, dema.herzallah@rockets.utoledo.edu, jennifer.peterson@utoledo.edu, caroline.menezes@utoledo.edu

Introduction. The prevalence of Parkinson's disease (PD) globally is increasing rapidly and might even be the fastest among the neurodegenerative disorders (Bloem *et al.*, 2021). PD is characterized by both motor and nonmotor features with cardinal signs including bradykinesia, resting tremors, rigidity (cogwheel or lead pipe rigidity) and postural instability (Jankovic, 2008). PD symptoms affect both voice and handwriting (Thomas et al., 2017) with 5% of the population displaying micrographia (McLennan et al., 1972) even before onset of the motor symptoms. Micrographia is an impairment of fine motor skill that manifests as reduced amplitude of the strokes in handwriting or as a progressive reduction of strokes (Kanno et al., 2019). Additionally, handwriting strokes get smaller as processing demands increase, such as when dual tasks are required (van Gemmert et al., 1999). Micrographia and hypophonia are highly correlated in Parkinson's disease (McLennan, et al., 1972; Wagle Shukla et al., 2012). Hypophonia is reduced amplitude of voice resulting in soft voice. Interestingly, people with PD often judge their speech to be loud indicating abnormalities in higher-order sensorimotor integration. Both micrographia and hypophonia appear to accompany bradykinesia, which is slowness of movement. Taken together, these data indicate a potential overlap in these pathophysiological responses (Murray et al., 2000). Research shows that there is a tight link between the planning of speech and hand movements in healthy people (Vainio *et al*., 2014; Salmelin & Sams 2002; Gentilucci *et al.,* 2001) and that systems governing speech and gesture are tightly linked in the mutual cognitive activity of language (Iverson and Thelen 1999; Gentilucci, *et al.,* 2001; Grossi, Maitra, & Rice, 2007). In PD, the work of Schneider *et al*., 1986, 1987 (as reported in Ho *et al*., 2000) has found both sensorimotor integration and proprioceptive abnormalities in the orofacial, hand and arm region of the brain, making it difficult for patients to use sensory information to complete a motor act. Speech therapy in PD patients focuses on speak with intent and loudness to address bradykinesia. We performed a preliminary study to investigate the relationship between the dual tasks of speaking and writing, before and after speech therapy, focusing on changes occurring when participants were asked to speak with a soft or loud voice while performing handwriting.

Methods. Handwriting and speech samples were collected from five patients who participated in the Parkinson's Speech Clinic summer of 2023. One was eliminated from this preliminary study due to further neurological diagnosis nonindicative of Parkinson's, and the other due to high cognitive decline that resulted inability to follow the directions, resulting in 3 participants in the study Handwriting samples were collected pre and post speech therapy intervention, and all patients were on their prescribed medication at the time of data collection. All patients received the SPEAK OUT! therapy protocol where they were trained to speak with intent. To test what effect speech has on handwriting, participants were instructed to write a series of 2-letter syllables, words and their name. This study analyzed the syllables (ha, li, and ti) only, on a letter sized unlined white paper using standardized pen. The syllable tasks were repeated 5 times per item. While writing the target item they were asked to enunciate syllables in their normal voice, loud voice, and soft voice (dual task). Participants were also asked to write syllables without voicing and voice without writing (single tasks). Instructions for loud and soft voice were provided in a randomized block design., Participants were given breaks if they required. No other instructions such as use of lower case or cursive styles were given to the participants. As a result, pre and post writing samples were different for some participants. To analyze the handwriting the recorded items were magnified to 400 for enhanced measurement precision. Subsequently, the largest stroke of the syllable was measured as the maximum height for that given syllable and the smallest stroke was measured as the minimum height. The initial and terminal point of the syllable were measured to determine horizontal syllable length, a. Due to extreme variation from individual participants and between the pre and posttest these values were normalized by calculating the area of a trapezoid where the maximum height and the minimum height formed the sides of the trapezoid and the height of the trapezoid was the horizontal length. The calculated area was then used for the analyses instead of individual stroke lengths.
During this task all audio outputs were recorded using a steady state Marantz portable recorder and head worn microphones. All audio files were then parsed and labeled using Praat (Boersma & Weenink (1992–2022). Duration and intensity measurements were made for each target item and averaged over the repetitions.

Results. The average duration of the acoustic syllable was calculated for all syllables separated by voice conditions and pre and post speech therapy. Similarly, average trapezoidal area for the written syllables was calculated. Results are

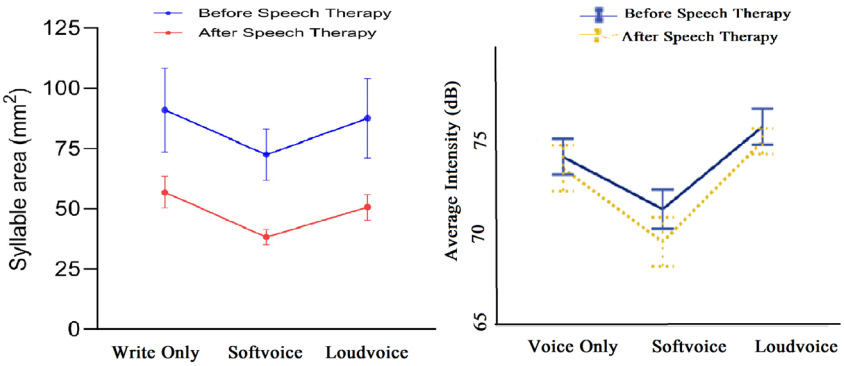


Figure 1: *Line graphs depicting changes in syllable duration and handwriting for the conditions of voice only, handwriting only, soft voice and loud voice separated for pre and post speech therapy.*

depicted in Figure 1. This figure shows that handwriting was greatly reduced following therapy but not much difference was observed in speech. The biggest difference in speech was an increase in speech range following therapy due to the lower values achieved in soft voice. However, conversation level loudness is 65 dB, and we see that speech recorded both pre and post was above conversation level. Participants handwriting and voice were distinctly different for the different conditions measured here. Soft voice condition revealed a loudness level lower than normal with a corresponding reduction in handwriting, while loud voice condition resulted in increased loudness level and increased handwriting. Results of average acoustic syllable durations indicated longer durations for post therapy syllables when compared to pre therapy syllables. It was significantly longest in the loud voice condition, with the largest variability evident in the soft voice condition. More details will be provided in the following paper.

Discussion. It is not clear why handwriting area decreased following therapy, but some understanding might be gleaned from the behavior of voice and handwriting between the dual task of loud voice and the single task of voice or writing alone. In the single task, both voice and handwriting were relatively good, but in the dual task, voice goals were prioritized over handwriting confirming the findings of van Gemmert (1999). Further, observations of the raw data also showed evidence where the syllable was voiced more times than written down. Handwriting and voicing were not synchronously produced, with voice production often leading handwriting. As voice amplitude increased, speech rate also increased but handwriting area decreased. Further analyses need to be conducted to determine if decreased handwriting area indicates a more controlled writing by comparing variation in stroke sizes between the different conditions. Finally, more data needs to be collected to generalize these findings to the symptomatology of PD.

References

Bloem, B, Okun, M.S. & Klien, C. (2021). Parkinson's Disease. *The Lancet*, v 397, 2284-2303.

Boersma, P & Weenink, D. (1992–2022) Praat: doing phonetics by computer [Computer program].Version 6.2.06, retrieved 23 January 2022 from https://www.praat.org.

Gentilucci, M., Benuzzi, F., Gangitano, M. and Grimaldi, S. (2001). Grasping with hand and mouth: a kinematic study on healthy subjects. *J. Neurophysiology* 86, 1685-1699. Gentilucci, M. (2003). Object motor representation and language. *Experimental Brain Research, 153,* 260–265.

Grossi, J. A., Maitra, K. K., & Rice, M. S. (2007). Semantic priming of motor task performance in young adults: Implications for occupational therapy. *American Journal of Occupational Therapy, 61,* 311–320.

Ho, A. K., Bradshaw, J. L., & Iansek, T. (2000). Volume perception in parkinsonian speech. *Movement disorders : official journal of the Movement Disorder Society*, *15*(6), 1125–1131. https://doi.org/10.1002/1531-8257(200011)15:6<1125::aid-mds1010>3.0.co;2-r

Iverson, J. M. and E. Thelen (1999). "Hand, mouth and brain. The dynamic emergence of speech and gesture." Journal of Consciousness studies **6**(11-12): 19-40.

Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008;79:368–376.

Kanno, S., Shinohara, M., Kanno, K., Uchiyama, M., Nishio, Y., Baba, T., Takeda, Y., Fukuda, H., Mori, E. & Suzuki, K. (2019). Neural substrates underlying progressive micrographia in Parkinson's disease. Brain Behav. 2020;10:e01669. https://doi.org/10.1002/brb3.1669.

McLennan, J. E., Nakano, K., Tyler, H. R., & Schwab, R. S. (1972). Micrographia in Parkinson's disease. Journal of Neurological Sciences, 15, 141–152. https://doi.org/10.1016/0022-510x(72)90002-0

Murray BJ, Llinas R, Caplan LR, et al. Cerebral deep venous thrombosis presenting as acute micrographia and hypophonia. Neurology 2000;54:751e3.

Salmelin, R. and Sams, M. (2002). Motor cortex involvement during verbal versus non-verbal lip and tongue movements. *Human Brain Mapping*, 16, 81-91.

Vainio, L., Tiainen, M., Tiippana, K. & Vainio, M. (2014). Shared processing of planning articulatory gestures and grasping. *Experimental Brain Research*, 232: 2359-2368.

Van Gemmert AW, Teulings HL, Contreras-Vidal JL, Stelmach GE. (1999). Parkinson's disease and the control of size and speed in handwriting. Neuropsychologia;37:685–694.

Wagle Shukla, A., Ounpraseuth, S., Okun, M.S., Gray, V., Schwankhaus, J., & Steven Metzer, W. (2012). Micrographia and related deficits in Parkinson's disease: a cross-sectional study. *BMJ Open* 2012;2: e000628. doi:10.1136/ bmjopen-2011-000628.

Progressive speech type adaptation to distance: frequencies of /a/ as a case study

Julien Meyer[1], Adèle Denis[2], Laure Dentel[3]

[1]Université Grenoble Alpes, CNRS, Gipsa-lab, France
[2]Université de Rennes, France
[3]The World Whistles Research Association, France

julien.meyer@cnrs.fr , adele.denis@etudiant.univ-rennes.fr,
ldentel.lab@gmail.com

Introduction. Human languages can be encoded and decoded by speakers and listeners with a certain amount of adaptability and flexibility. Such human adaptive abilities in speech communication have enabled the development of different natural speech types (e.g. whispering, shouting..). Some of these speech types are efficient in situations of distance communication and the present study deals with three of these: strong speech, shouted speech and a more extreme transformation called whistled speech (Meyer 2020). In whistled speech, people articulate words while whistling and thereby transform spoken utterances by simplifying them into whistled melodies (Busnel & Classe 1976). Speakers of non-tonal languages (such as Spanish, Greek, Wayãpi…) transpose linguistic segments – typically vowels and consonants - into whistled pitches, lending alternative insights into how the phonetic expression of phonemes can be drastically reduced without hindering cognitive reconstruction. Typically, the vowels are emitted at different whistled pitch levels depending on the frequency distribution of different spoken vowel qualities because whistlers approximate the spoken articulatory movements to pronounce words while whistling. Because whistling occurs in the front oral cavity, it very often resembles spoken formant 2 (Rialland 2005). In voiced speech and in whistled speech, speakers tacitly adapt to constraints of distance and noise to transmit their message to the receiver (Fux 2012; Zahorik & Kelly 2017; Meyer 2015; Meyer *et al.* 2018). In the present study, we present an original protocol designed to explore this adaptation in ecologically valid contexts, by increasing progressively the distance to which the speakers had to project their sentences. We also present the first results of measures on the frequencies of spoken, shouted and whistled vowels /a/ of the corpus.

Methods. The main challenge was to develop a same experimental protocol and recording setting for three different speech forms targeting largely different distances. Moreover, in order to approach realistic conditions of communication for the participants, we gave priority to short sentences - with an average of 8.3 syllables and 4.8 words - commonly used in rural settings [for example: "Nos vemos en mi casa" (eng: *we meet at my house*), "¿Antonio, vienes a cenar?" (eng: *Antonio, do you come for dinner?*)]. We recorded a corpus of 30 sentences with three native Spanish speakers from Tenerife Island. All participants were long term (>5 years) teachers of whistled speech and thus experts in whistled Spanish. The recordings were made in an open field characterized by very low - assumed to be negligible - reverberation indices. Moreover, background noise was checked as below 40 dB(A), and wind below 2m/s throughout the measured data. The task of the speakers was to target specific distances while pronouncing each sentence with the voice (targets were at 7m for speaking and 70m for shouting) and also with whistling (targets were at 70m, 140m, 500m). Speakers had the task to make as if they had to transmit the sentence to an imaginary listener situated at these distances (localized precisely by a tree or a small building). For voiced speech we never asked them to shout but chose the distance of 70 m so that shouting would be attained (see for example Meyer *et al.* 2018, and we additionally checked that mean sound pressure level of vowels was maintained >80 dB(A) at 1m from the mouth). Dual recording was made at 7m (Zoom H4n and Rion NL42 Sonometers, using built–in microphones) while an additional control recording was made at 70m (Zoom H4n with same level as the one at 7m). For this study we focused on recordings made with the Zoom H4n at 7m in order to measure values in production rather near to the speakers. Annotations and segmentations were double-checked by two experimenters, keeping only the clearly pronounced vowels (pauses were made to enable the speakers to rest and drink, but some whistled productions were still altered by the efforts) and the vowel nuclei, or middle portion of the vowel which is the less altered by coarticulations (Busnel & Classe 1976, Rialland 2005, Meyer 2015). For voiced speech (strong spoken and shouted), we measured the F0 of each extracted /a/ vowel, as well as the Formant 2 values. For whistled speech, we measured the F0 of each extracted /a/ whistled vowel.

Results. We analyzed separately voiced and whistled vowels given their different nature in frequency. Results show significant increase in voice F0 between strong and shouted speech (from 175 Hz for 7m to 295 Hz for 70 m) but no significant difference in voice Formant 2, for all and each speaker(s) (see examples of recordings and results for F0 in one whistler in Figure 1). This was verified by running Generalized Mixed Models on these Freq. variables with Distances as a fixed effect (and while comparing with other models adding Speakers as fixed effects and/or Syllables types as random effects). For whistled speech, the same type of GLM showed differences in progressive adaptation between

speakers as one of them produced highly different frequencies at the three distances (speaker of Figure 1), while another showed more frequency proximity between 140 and 500m, whereas the third one clearly grouped /a/ whistled frequencies between 70m and 140m.

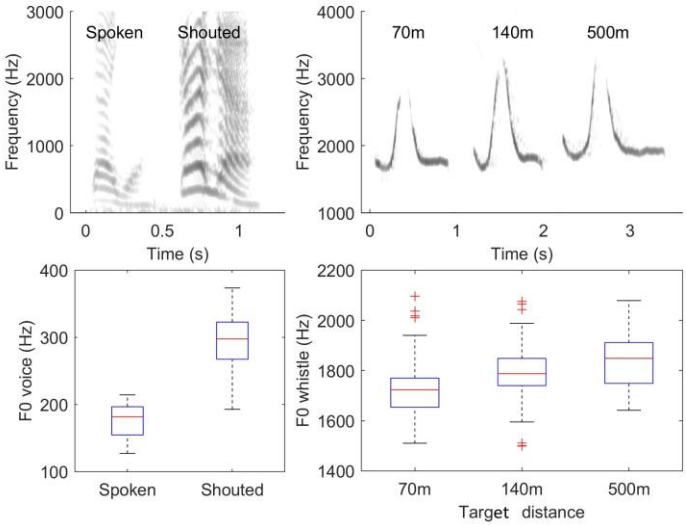


Figure 1: *Spoken, shouted (above left), and three whistled forms (above right) of the Spanish word /kaja/. Below are the pitch values for each speech type of all /a/ vowels for one of the three speakers (sentence corpus).*

Discussion. These results confirm the observations already made previously concerning voiced F0 increase as vocal effort increases, and as concomitant to a tacit Lombard effect during speech production for distance communication (e.g. Fux 2012; Zahorik & Kelly 2017; Meyer *et al.* 2018). They also show the relative stability of mean values of formants, even if a wider dispersion of pitch values was found for shouting than for strong voice. Further studies should explore other vowels and other formants, as well as the relations of proximities between formants as this appears to be an important factor influencing vowel perception and their imitation into whistles (Meyer 2015). The results on whistling show that even if whistlers transpose segmental characteristics of vowel qualities of /a/, they increase frequency as they project farther their whistle. It is possible that this is a consequence of an increased amplitude of production rather than an additional adaptation in frequency to environmental constraints. This point should be further explored. Moreover, different whistlers had different thresholds to jump to higher frequencies, showing that exploring inter-whistler variability in progressive adaptation to distance is an interesting perspective.

References

Busnel, R-G., & Classe, A. (1976). Whistled languages. Berlin Heidelberg, Springer.

Fux, T. (2012). Vers un système indiquant la distance d'un locuteur par transformation de sa voix. Ph.D. Dissertation. Grenoble : Université de Grenoble.

Meyer J., Meunier F., Dentel L., Do Carmo Blanco N., & Sèbe F. (2018) Loud and Shouted Speech Perception at Variable Distances in a Forest. Proceedings of Interspeech 2018, Hyperabad, India. 2285-2289

Meyer J (2020) Coding Human Languages for Long-Range Communication in Natural Ecological Environments: Shouting, Whistling, and Drumming. In: Aubin T., Mathevon N. (eds) Coding Strategies in Vertebrate Acoustic Communication. Animal Signals and Communication, vol 7. pp.91-113, Springer.

Meyer, J. 2015. Whistled languages. A Worldwide Inquiry about human whistled speech. Berlin, Springer

Rialland A. (2005). Phonological and phonetic aspects of whistled languages. Phonology 22:237–71.

Zahorik & Kelly, J.W. (2007). Accurate vocal compensation for sound intensity loss with increasing distance in natural environments. Journal of the Acoustical Society of America Express Letters, 122 (5) 143–150.

.

Predicting Articulatory Landmarks with Critically-Damped Oscillators and General Tau Theory

Christopher Geissler[1], Jyothiraditya Nellakra[1]

[1]*Carleton College*

cgeissler@carleton.edu, neallakraj@carleton.edu

Introduction. The mathematics of dynamical systems has proven to be a fruitful way to relate continuous and discrete properties of speech (Iskarous 2017; Mücke, Hermes, and Tilsen 2020). In this paper, we compare the ability of two models, critically-damped oscillators and General Tau Theory, to predict individual points in kinematic data.

Articulatory movements have been modeled as critically-damped mass-spring oscillators by Saltzman and Munhall (1989) in Task Dynamics. Among the benefits of this approach is the ability to describe intergestural timing in terms of phase, and to coordinate gestures by coupling the oscillators, as in Nam and Saltzman (2003).

More recently, Elie, Lee, and Turk (2023) have applied General Tau Theory to speech. This model, adapted from work on non-speech motor control, is based on the time-to-closure of "gaps" rather than mass-spring systems. Elie, Lee, and Turk (2023) found that a Tau-based approach compared favorably to coupled-oscillator implementations when fitting kinematic data. That study globally compared the fit of several models to a corpus of electromagnetic articulography (EMA) data of English speech.

The present study instead focuses on experimental stimuli collected to study gestural timing, and uses a typologically-different language, Tibetan. We test coupled-oscillator and General Tau models by fitting each to articulatory trajectories, then comparing their predictions for specific points that are commonly used as landmarks for characterizing articulatory gestures. Our findings highlight advantages of each model, and demonstrate how differences in the curves translate to differences at salient kinematic landmarks.

Methods. Predictions of the two models–critically-damped oscillators and General Tau Theory–were compared with each other and with original kinematic data. The data consisted of electromagnetic articulography recordings collected as part of Geissler (2021). Six speakers (four female) of diaspora Tibetan were recorded producing one- and two-syllable Tibetan words in a carrier phrase. The target syllables consisted of /m/ followed by the back vowels /u o a/ and either high or low tone. Syllables with and without coda consonants were included, and the /mV/ sequence was always word-initial. Lip aperture was used for the consonantal gesture, and tongue dorsum retraction for the vowel gesture. Gestural landmarks were calculated in *Mview* (Tiede 2005): position and velocity were recorded at the point of peak velocity toward target, the points where 20% of peak velocity was reached during acceleration and deceleration, and at the point of maximum constriction. These landmarks were recorded both in the closing and opening portions of each movement.

Parameters for each model were set using certain landmarks, then used to predict the spatio-temporal coordinates at other landmarks. For the coupled oscillator model, the peak velocity and position at peak velocity were used to calculate the natural frequency of the oscillator. For the Tau model, $\kappa = 0.4$ was used following the results of Elie, Lee, and Turk (2023), and the magnitude and duration of gestures were taken as the difference between kinematically-identified gestural onset and target attainment. Both models were then used to calculate predicted positions for the timestamps identified in the kinematics for peak-velocity and target attainment.

Results. A comparison of predicted with actual data is presented in **Figure 1**. Analysis and model comparison with linear mixed-effects models confirmed that interactions between landmark and model/data type were significant for both time and distance.

As compared to results from kinematic thresholds, the critically-damped oscillator model tended to predict that landmarks would take place earlier in time and closer to the target. The General Tau model generally predicted that landmarks would take place later and closer to the target. These patterns broadly held for both closure and release landmarks, and for both the consonantal lip gesture and the vocalic tongue dorsum gesture.

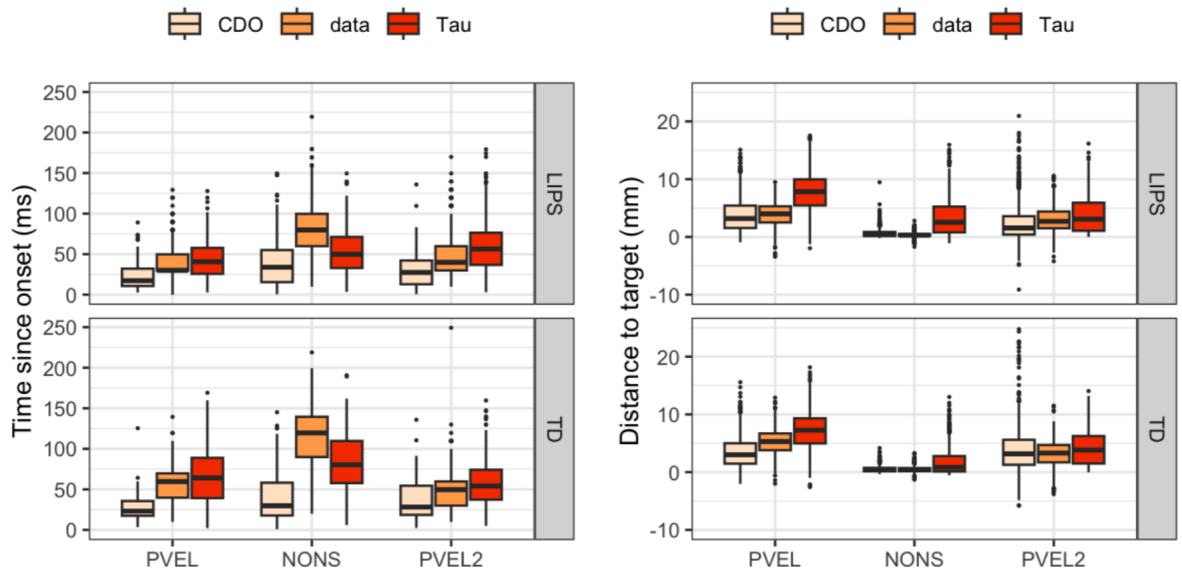


Figure 1: *Tibetan /mV/ sequences. CDO = critically-damped oscillator; data = kinematically-defined landmarks; Tau = General Tau model. PVEL/PVEL2 = point of peak velocity toward/away from target; NONS = (gestural) nucleus onset*

Discussion. These results highlight the differences in the trajectory shape generated by each model. Critically-damped oscillators move rapidly, then slow to asymptotically approach the target; Tau-derived trajectories unfold gradually and (for the right value of κ) symmetrically, and reach the target at a known point.

Constructing these models also called attention to the importance of careful definitions for the start and end of a gesture. Both oscillator and Tau models required kinematic landmarks: the oscillator model used the onset of the gesture (along with the peak velocity), while the Tau model used both beginning and end of each gesture. Using different values, such as the point of maximum constriction rather than nuclear onset for the Tau model, leads to different results. Careful consideration for the use of particular landmarks is crucial to accurately comparing models.

This study was limited by the range of materials and the relatively simple versions of the models used. For example, we would expect to find better-fitting curves had the oscillator model used gradient activation like that of Sorensen and Gafos (2016). Nevertheless, the results demonstrate that generating predictions for specific points allows for models to be tested against each other and against speech data.

References.

Elie, Benjamin, David N. Lee, and Alice Turk (June 2023). "Modeling trajectories of human speech articulators using general Tau theory". en. In: *Speech Communication* 151, pp. 24–38. DOI: 10.1016/j.specom.2023.04.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167639323000614 (visited on 06/19/2023).

Geissler, Christopher (2021). "Temporal articulatory stability, phonological variation, and lexical contrast preservation in diaspora Tibetan". PhD thesis. Yale University. URL: https://elischolar.library.yale.edu/gsas_dissertations/52.

Iskarous, Khalil (Sept. 2017). "The relation between the continuous and the discrete: A note on the first principles of speech dynamics". en. In: *Journal of Phonetics* 64, pp. 8–20. DOI: 10.1016/j.wocn.2017.05.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0095447017301006 (visited on 10/17/2020).

Mücke, Doris, Anne Hermes, and Sam Tilsen (Feb. 2020). "Incongruencies between phonological theory and phonetic measurement". en. In: *Phonology* 37.1, pp. 133–170. DOI: 10.1017/S0952675720000068. URL: https://www.cambridge.org/core/product/identifier/S0952675720000068/type/journal_article (visited on 07/20/2020).

Nam, Hosung and Elliot Saltzman (2003). "A competitive, coupled oscillator model of syllable structure". In: *Proceedings of the 15th International Congress of the Phonetic Sciences*.

Saltzman, Elliot and Kevin Munhall (Dec. 1989). "A Dynamical Approach to Gestural Patterning in Speech Production". en. In: *Ecological Psychology* 1.4, pp. 333–382.

Sorensen, Tanner and Adamantios Gafos (Oct. 2016). "The Gesture as an Autonomous Nonlinear Dynamical System". en. In: *Ecological Psychology* 28.4, pp. 188–215. DOI: 10.1080/10407413.2016.1230368. URL: https://www.tandfonline.com/doi/full/10.1080/10407413.2016.1230368 (visited on 12/15/2023).

Tiede, Mark (2005). *Mview: software for visualization and analysis of concurrently recorded movement data.*

Exploring Vowel Reduction in French Casual Conversations

Kübra Bodur[1], Corinne Fredouille[2], Christine Meunier[1]

[1]Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[2]Avignon Université, CNRS, LIA, Avignon, France
kubra.bodur@univ-amu.fr

Introduction. Vowel reduction in speech production is a complex phenomenon of variation with implications across languages (Meunier et al., 2006; Gendrot & Adda-Decker, 2007). This phenomenon refers to the alteration of vowel quality, typically involving the reduction of a vowel's full articulatory qualities towards a more centralized, less distinct sound, often occurring in unstressed syllables. Previous research on reduction in French (Gendrot & Adda-Decker, 2005, 2007) demonstrated that vowels with shorter/reduced durations have more central spectral values. The intricate interplay among contextual, sociolinguistic, cognitive, and physiological factors influences vowel articulation, thereby shaping the acoustic space occupied by vowels in speech. Therefore, understanding the relationship between vowel characteristics and other forms of reduction in speech is pivotal in comprehending the dynamics of informal spoken language.

Methods. Our study aims to explore vowel characteristics and reduction in a corpus of casual French conversations (*CID,* (Bertrand et al., 2008)) between 8 pairs of colleagues, each of which lasting around one hour. Extracting the vowels (/a/, /e/, /i/, /o/, /ø/, / u/, /y/) produced by 16 speakers (*10 F* and *6 M*), we obtained 1/ Vowel Space Areas (VSA, hereafter) (Chung, 2012) by computing the area of the F2 x F1 vowel space as a polygon connecting formant means (pVSA), 2/ the Vowel Distinctivity Index (VDI, hereafter) (Huet, 2000; Meunier & Ghio, 2018) to quantify vowel distinctiveness of a speaker based on the F1 and F2 values. A second focus of this study is to establish a relationship between these speaker-specific metrics and the frequency of reductions they produced. Our hypothesis posits that speakers with smaller pVSA and VDI values are inclined to produce more lexicalized *(e.g.,* ʃɛpa instead of "je ne sais pas", Bodur et al., under revision) and non-lexicalized reductions compared to those with higher metric values.

Results. The variability in VDI values and VSA sizes was considerable among speakers in conversations. Figure 1 illustrates individual VSA examples for two speakers, while Figure 2 displays reduction counts and VDI values. A weak negative correlation exists between VDI and non-lexicalized reductions (*-0.059, p=0.8286*), whereas a moderate negative correlation is observed between VDI and lexicalized reductions (*-0.353, p=0.1797*), suggesting an increase in reduction ratios as VDI decreases. Moderate correlations are found between pVSA and reduction ratios. Negative correlation with both lexicalized (*-0.44, p= 0.08818*) and non-lexicalized reductions (*-0.42, p=0.1047)* indicate trends, albeit inconclusive.

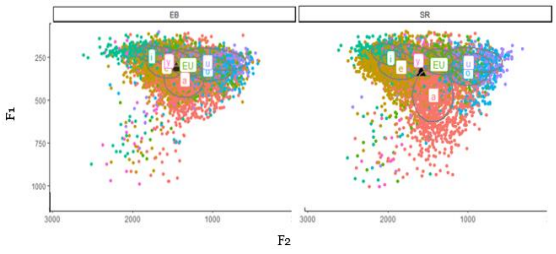


Figure 1: The Vowel Space Areas for two male speakers. The centers of gravity are marked by black triangles. (EB has a smaller VSA and less distinctive vowels while SR has a bigger VSA and a higher VDI).

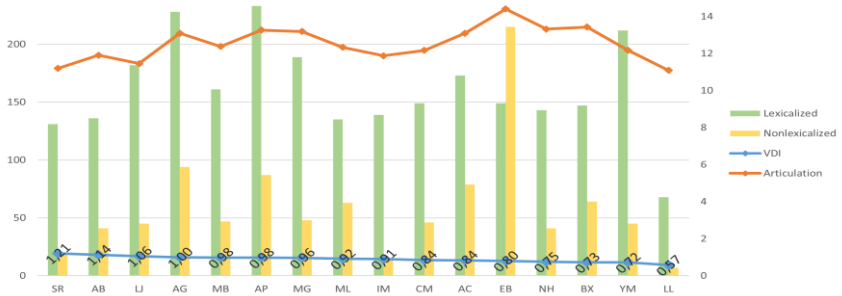


Figure 2. VDI and articulation rates across speakers, using bars for lexicalized (green) and non-lexicalized (yellow) reductions, and lines for VDI (blue) and articulation rates (orange).

Further investigation revealed a weak negative correlation between VDI and articulation rates (*-0.187, p= 0.4882*), suggesting that as the articulation rate increases, VDI might tend to decrease (Figure 2). However, this effect was not statistically significant.

For a subset of speakers, we computed VDI across different speaking conditions (isolated vowels, monosyllabic words and texts made out of monosyllabic words) in addition to the conversational context. Preliminary findings indicate significantly reduced spectral qualities of vowels in conversational speech (*p<0.001* for vowel-conversation contrast, *p=0.00518* for word-conversation contrast), highlighting the impact of speaking situations on vowel characteristics (Smiljanić & Bradlow, 2009; Gendrot et al., 2012). Figure 3 represents VDI values for these speakers in various conditions. Interestingly, while inter-speaker variability in VDI was significant in conversations, we observed a greater intra-speaker variability across different speaking contexts; VDI decreased gradually as the context expanded.

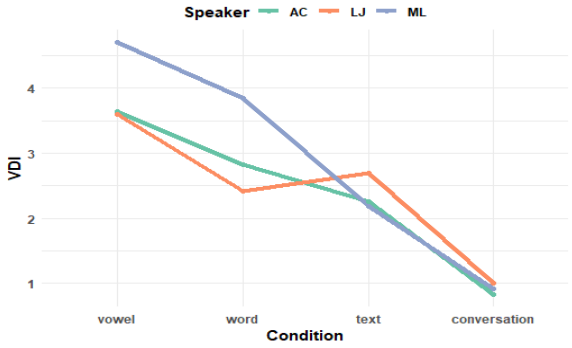


Figure 3. VDI values calculated for the vowels produced in different conditions by three speakers of the corpus.

Discussion. By examining vowel reduction within casual French conversations, our study reveals significant inter and intra-speaker variability in VDI and pVSA values. Vowels in casual speech exhibit a more reduced articulation, characterized by more centralized spectral values, highlighting acoustic alterations among speakers. Moreover, the variation due to speaking conditions goes beyond the inter-speaker variation. The complexity of speech production emerges from the intricate relationship between vowel characteristics and spoken language reduction. Correlations observed between reduction instances and acoustic measures illustrate this nuanced relationship. While vowel reduction is inherent in conversational speech, our results suggest that reduction in vowel quality alone might not fully explain the occurrence of lexicalized and non-lexicalized reductions. Understanding this multifaceted interplay between acoustic measures and the diverse spectrum of reduction instances to highlight the intricate dynamics influencing speech patterns in conversations is thus crucial. Further exploration is essential to comprehensively grasp the multifaceted nature of vowel reduction and its interaction with conversational phenomena.

References

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID: Corpus of Interactional Data. Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique des Langues*, 49(3).

Chung, H., Kong, E. J., Edwards, J., Weismer, G., Fourakis, M., & Hwang, Y. (2012). Cross-linguistic studies of children's and adults' vowel spaces. *The Journal of the Acoustical Society of America*, *131*(1), 442-454.

Gendrot, C., & Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: An automatic analysis of large broadcast news corpora in French and German. In *Proceedings of Interspeech conference* (pp. 2453–2456), Lisbon, Portugal.

Gendrot, C., & Adda-Decker, M. (2007). Impact of duration and vowel inventory size on formant values of oral vowels: an automated formant analysis from eight languages. In *Proceedings of the 16th international congress of phonetic sciences* (pp. 1417-1420). Saarbrüken, Germany.

Gendrot, C., Adda-Decker, M., & Schmid, C. (2012). Comparaison de parole journalistique et de parole spontanée : analyses de séquences entre pauses (Comparison of journalistic and spontaneous speech: analysis of sequences between pauses)[in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP* (pp. 649-656).

Huet, K., & Harmegnies, B. (2000). Contribution à la quantification du degré d'organisation des systèmes vocaliques. *Actes des JEP'2000*, 225-228.

Meunier, C., Espesser, R., & Frenck-Mestre, C. (2006). Phonetic variability as a static/dynamic process in speech communication: a cross linguistic study. In *Laboratory Phonology (LabPhon)* (pp. pp-129).

Meunier, C., & Ghio, A. (2018). Caractériser la distinctivité du système vocalique des locuteurs. In *XXXIIe Journées d'Études sur la Parole* (pp. 469-477).

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and linguistics compass*, *3*(1), 236-264.

Speaking style influence on vowel length opposition in Jordanian Arabic

Mohammad, Abuoudeh[1] Jalal, Al-Tamimi[2] Olivier, Crouzet[3]
(1) Department of Language and Linguistics, Al-Hussein Bin Talal University, Ma'an, Jordan
(2) Université Paris Cité, CNRS, Laboratoire de linguistique formelle, F-75013 Paris, France
(3) Laboratoire de Linguistique de Nantes (LLING), UMR6310–Université de Nantes/CNRS, France
mohammad.a.abuoudeh@ahu.edu.jo | jalal.al-tamimi@u-paris.fr | olivier.crouzet@univ-nantes.fr

Introduction Changing speaking style can provoke temporal and spectral variations of the produced segments (Lindblom and Lindgren, 1985). These variations take place due to the change in strategies of speech production. Some speech situations must be realized with a high degree of perceptual contrast; others require less and allow more variability. Consequently, the acoustic properties of the same sound show a wide range of variations reflected along a continuum varying from hypo- to hyper-articulation (Farnetani and Recasens, 2010; Lindblom, 1990). The present study aims to examine the impact of speaking style on vowel spectral and temporal information in a context where phonologically long and short vowels are opposed. Many studies investigated the influence of changing the speaking style on vowel quality and quantity in many languages (Blaauw, 1992; Bolotova, 2003; DiCanio et al., 2015; Meunier and Espesser, 2011). The common point of these studies is that in spontaneous/casual speech, segment duration and vowel space are reduced compared with read/clear speech. Few studies examined the relationship between long and short vowels when speaking style changes. For example, DiCanio and Whalen (2015) found an asymmetrical influence of speaking style on long and short vowels in Arapaho. Long vowel duration is more influenced by changing speaking style, while its vowel space is less impacted by this factor in comparison with short vowels. Similar results were found in English tense-lax opposition where the duration of tense vowels is more impacted than the duration of lax vowels due to speaking style variation. In addition, the latter has fewer consequences on vowel space of lax than tense vowels.
The purpose of this research is to examine to which extent variations from story reading to storytelling would influence the durational and spectral information for long and short vowels in Jordanian Arabic (JA). JA contains 3 short vowels and their long counterparts /i, iː, a, aː, u, uː/ in addition to 2 other long vowels /eː, oː/. The importance of vowel duration in JA depends on the vowel timbre; /a, aː/ are mainly differentiated by duration, /u, uː/ are distinguished by both duration and spectral information, and /i, iː/ are mainly distinguished by spectral information (Abuoudeh, 2018; Al-Tamimi, 2007). Following the previous studies, we are expecting that the story reading style would lead to a longer vowel duration and a larger vowel space compared with the storytelling style. In addition, this influence would be asymmetrical between long and short vowels.
Methods 10 Jordanian speakers (5 females and 5 males), were asked to read Little Red Riding Hood story from a text on a computer screen and then they were asked to tell the same story without the text[1]. Both tasks (reading and telling) were transcribed, segmented and forced-aligned using the online `Arabic WebMAUS Basic` service (Al-Tamimi et al.; Kisler et al., 2017). The results of the forced alignment were corrected manually afterward using `Praat` software. Segment duration, F1, F2, F3, and f0 were automatically extracted using a `Praat` script and were saved in a *.csv* file. The frequencies of F1 and F2 of all speakers were normalized using Lobanov method. Data analysis was performed using the `R` program.
Results All speakers produced 4972 vowels in reading task and 3992 vowels in telling task as detailed in Table 1. It was expected to have less realization in the telling task than in the reading task because the reader would omit some events or phrases while he or she was telling the story. Descriptive analysis indicates that the two studied speaking styles

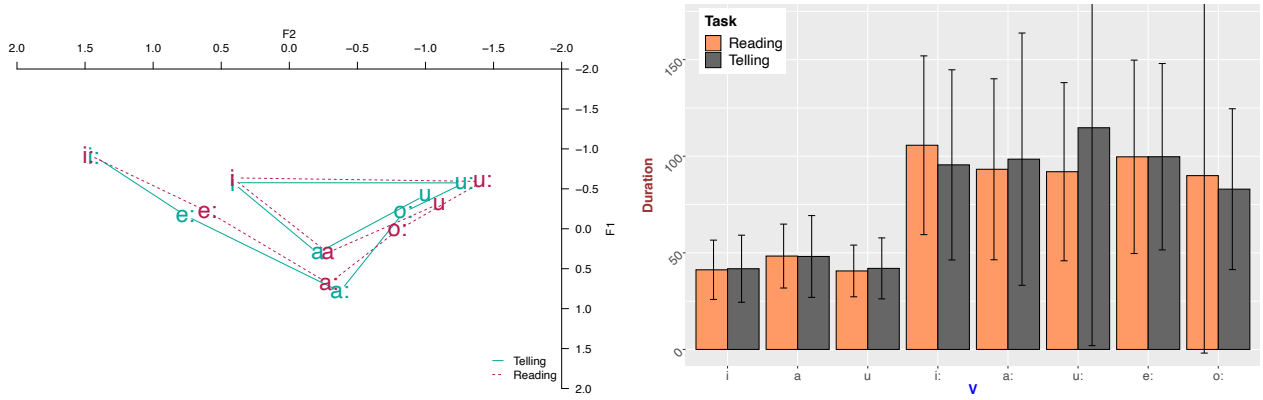
	i	iː	a	aː	u	uː	eː	oː
reading	1120	393	1664	1185	81	188	278	63
telling	942	360	1211	871	155	182	180	91

Table 1: Number of realisations of each vowel in each speaking style.

have a small impact on vowel duration and vowel space (Figure 1). This observation is confirmed by linear mixed model analysis[2] that shows no significant differences between reading and telling tasks for duration ($F_{(1,7)} = 0.30, p = .587$),

[1]The present data is part of a speech database under construction of Jordanian Arabic (SDJAD) composed of over 100 participants from various regions
[2]p-values are derived using the Satterthwaite approximation.

F1 ($F_{(1,7)} = 0.48, p = .494$), and F2 ($F_{(1,7)} = 0.0001, p = .99$). Theses results indicate that the vowel length opposition



(a) F1 F2 in Lobanov

(b) Duration in ms

Figure 1: Vowel space (a) and vowel duration (b) in function of task.

in JA is not influenced by changing the speaking style from telling to reading.

Discussion This study aimed at evaluating the impact of changing speaking style on vowel opposition in JA. No impact was observed in our data. Consequently, this is not in line with the previous studies mentioned above (among others) describe that going from clear to casual speech leads to temporal and spectral variations. These findings could be explained by the fact that these two speaking styles are so close to each other, in a language which has a phonemic length contrast. In addition, the importance of duration separation between long and short vowels in JA would diminish the temporal effect and, therefore, the vowel space variation in these speaking styles. Interestingly, the qualitative differences across short and long vowels are still preserved. Studying other tasks from the SDJAD project (like the word in isolation, conversational speech, and image description) that are in process would verify the latter assumption.

References

M. Abuoudeh. *De l'impact des variations temporelles sur les transitions formantiques*. PhD thesis, Université de Nantes, 2018.

J. Al-Tamimi, F. Schiel, G. Khattab, N. Sokhey, D. Amazouz, A. Dallak, and H. Moussa. A Romanization System and WebMAUS Aligner for Arabic Varieties. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), © European Language Resources Association (ELRA), Licensed under CC-BY-NC-4.0*, pages 7269–7276.

J.-E. Al-Tamimi. *Indices dynamiques et perception des voyelles : Étude translinguistique en arabe dialectal et en français*. Thèse de doctorat, Université Louis Lumière - Lyon 2, 2007.

E. Blaauw. Phonetic differences between read and spontaneous speech. In *II International Conference on Spoken Language Processing ICSLP*, 1992.

O. Bolotova. On some acoustic features of spontaneous speech and reading in russian (quantitative and qualitative comparison methods). In *15th International Congress of Phonetic Sciences (ICPhS-15)*, 2003.

C. DiCanio and D. Whalen. The interaction of vowel length and speech style in an arapaho speech corpus. In *The 18th International Congress of the Phonetic Sciences*, 2015.

C. DiCanio, H. Nam, J. D. Amith, R. C. García, and D. H. Whalen. Vowel variability inelicited versus spontaneous speech: Evidence from mixtec. *Journal of Phonetics*, 48:45–59, 2015.

E. Farnetani and D. Recasens. Coarticulation and connected speech processes. In W. J. Hardcastle, J. Laver, and F. E. Gibbon, editors, *The Handbook of Phonetic Sciences*, pages 316–352. Wiley-Blackwell, second edition, 2010.

T. Kisler, U. Reichel, and F. Schiel. Multilingual processing of speech via web services. *Comput. Speech Lang.*, 45(C):326347, sep 2017. ISSN 0885-2308. doi: 10.1016/j.csl.2017.01.005. URL https://doi.org/10.1016/j.csl.2017.01.005.

B. Lindblom. Explaining phonetic variation : A sketch of H&H theory. In W. Hardcastle and A. Marchal, editors, *Speech production and speech modelling*, pages 403–439. Kluwer Academic Publishers, 1990.

B. Lindblom and R. Lindgren. Speaker-listener interaction and phonetic variation. *Phonetic Experimental Research at the Institute of Linguistics University of Stockholm-PERILUS*, 4:77–85, 1985.

C. Meunier and R. Espesser. Vowel reduction in conversational speech in french: The role of lexical factors. *Journal of Phonetics*, 39(3):271–278, 2011. ISSN 0095-4470. doi: https://doi.org/10.1016/j.wocn.2010.11.008. URL https://www.sciencedirect.com/science/article/pii/S0095447010000951.

Spatiotemporal variation of tongue dorsum characterizes the voicing contrast of American English bilabial coda obstruents

Daejin Kim

Department of Linguistics, University of New Mexico
daejinkim@unm.edu

Introduction. American English (AE) speakers distinguish the "voicing" properties of bilabial coda obstruent (C2) by producing lower *f0* and longer duration of the preceding vowel (V) (Maddieson, 1997). The tongue dorsum (TD), under the velum, has been increasingly understood as exhibiting articulatory properties that distinguish the voicing properties of bilabial obstruents in languages (Ahn, 2018; Coretta, 2020; Svirsky et al., 1997; Vazquez-Alvarez & Hewlett, 2007). This study argues that the voicing contrast of C2 should be evidenced in the spatiotemporal characteristics of TD, following the evidence of TD movement during the closure of obstruents. This secondary supralaryngeal movement, TD raising or lowering, was found to occur during the acoustic closure of onset /b/ increasing the subglottal air pressure before vocal folds vibration, which may result in lower *f0* of the following V. This study argues that it would also occur during the V to signal voicing. First, the movement characteristics of TD should correlate with the *f0* variation contrasting voicing according to the intrinsic characteristics of Vs. TD lowering, which may lower the larynx, is more likely to occur with low Vs and C2 /b/, resulting in lower *f0*, while TD raising, which may raise the larynx, is more likely to occur with high Vs with C2 /p/, resulting in higher *f0,* following the tongue-pulling hypothesis (Ohala, 1978). However, little evidence for such a pattern has been reported in AE. Moreover, AE is expected to show some TD movement differences in duration distinguishing C2 voicing. A study (Coretta, 2020) reported that tongue advancement duration is positively correlated with acoustic V duration in Italian and Polish, and this study examines whether AE speakers contrast consonantal voicing similarly. Taken together, this study hypothesizes that C2 /b/ should be associated with more frequent TD raising or lowering with longer distance and duration and greater intergestural timing between articulatory landmarks of C2 and V than C2 /p/. This would imply that TD movement distance and duration from the onset to the target and from the target to the offset may contribute to the acoustic characteristics of bilabial C2s, signaled by acoustic V duration.

Methods. Ultrasound tongue images synchronized with acoustic recording were collected from seven (three males and four females) native speakers of AE. Each speaker produced six monosyllabic target words (/hVC2/; V = /i, u, ɑ/, C2 = /p, b/) in phrase-medial position with broad (Q: What did you do {today, yesterday}? A: I wrote *a heap* (/hip/) on the {paper, note, board, letter}) and contrastive (Q: Did you write ***god*** on the {paper, note, board, letter}? A: No. I wrote a ***heap*** (/hip/) on the {paper, note, board, letter}) focus prominence six times. Tongue contours were estimated consecutively over time using DeepLabCut (Wrench & Balch-Tomes, 2022) by marking darker edges under the brighter reflections from the tongue surface. TD movements were annotated as the distance between the hyoid bone and the tongue surface under the velum when it starts (onset), reaches its target (target), and ends its movement (offset) for V and C2. Since not all tokens have visually apparent TD movement (Vazquez-Alvarez, 2007), tokens were annotated as 'Yes' if TD movements were visually evident or 'No' if not apparent in ultrasound tongue images. 'Yes' tokens were classified again based on TD onset-to-peak distance as either 'Raising' (≥ 0) or 'Lowering' (< 0). The conditional probabilities of the presence and absence of TD movement and the TD movement directions on bilabial C2s were estimated by building conditional inference trees (CITs) (Levshina, 2020) depending on consonants, vowels, focus prominence types, and individual speakers, using the *partykit* package (Hothorn et al., 2023) in *R*. TD movement differences among consonants were assessed with (i) onset-to-target and target-to-offset distance and duration, (ii) intergestural timing between V and C2 onsets, targets, and offsets, and (iii) timing of V and C2 onsets, targets, and offsets from the end of acoustic vowel duration. The statistical significances of movement variables were statistically tested as a function of *f0* peaks, vowel duration, consonants, vowels, and focus prominence types with random intercepts of speakers using the *lmer* package (Bates et al., 2015), followed by the Kenward-Roger post-hoc test using the *pbkrtest* package (Halekoh & Højsgaard, 2014) in *R*. Only statistically significant variables were included in the models.

Results. A CIT model estimating the factors that affect the presence and absence of TD movement (model accuracy rate = 0.79) found that C2 /b/ (probability of 'Yes = 0.93) is more likely to have TD constriction compared to C2 /p/ (0.64). Among /p/ tokens, /ɑ/ (0.89) is more likely to have TD constriction during V than /i/ and /u/ (0.56). In terms of TD movement direction, a CIT model (model accuracy rate = 0.76) estimated that /i/ and /u/ (probability of TD 'raising' = 0.82) are more likely to have TD raising movements than TD lowering movements compared to /ɑ/ (0.57). Regarding TD movements' characteristics, the models estimated that C2 /b/ has a longer movement distance (only when lowering) (β (difference estimate) = -0.6*) (**Figure 1 (a)**) and a longer onset-to-target duration (β = -6.4**) **(b)**, compared to C2 /p/. However, the deceleration distance (β = -0.4) and duration differences of TD (β = -3.1) were not statistically different between C2s. None of the other variables regarding TD movements, however, showed significant correlations with *f0*

peak, V duration, focus prominence, or V type in the models. Regarding intergestural timing, TD starts to move simultaneously for V away from C2 onsets between C2 and V (**(c)** β = -2.4). /b/ has earlier V targets (**(d)** β = -10.7*) and offsets of TD (**(e)** β = -12.6***) than /p/, resulting in less coarticulation between V and C2 in timing. If speakers shifted the TD timing of C2 with /b/, /b/ has later C2 target and offset of TD than /p/, resulting in articulatory expansion away from V. Relative to the acoustic end of V, /b/ has earlier onset, target, and offset of TD for both V (**(f)** V onset – V end: β = 39.7*; V target – V end: β = 40.3***; V offset – V end: β = 37.5***) and C2 (**(g)** C2 onset – V end: β = 35.9***; C2 target – V end: β = 29.8***; C2 offset – V end: β = 24.9***) compared to /p/. All TD movements of /b/ occur earlier than those of /p/, resulting in more temporal overlap with the V duration and TD movements for C2.

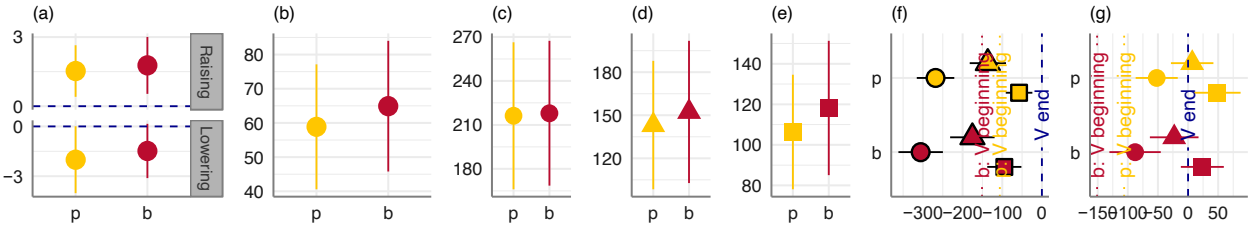


Figure 1: *(a)-(b) TD movement distance and duration for bilabial C2s. (c)-(e) Intergestural timing (.ms) between onsets (circles), targets (triangles), and offsets (squares) of V and C2. (f)-(g) Relative timing (.ms) of V and C2 onsets, targets, and offsets from the acoustic end of V (blue-dashed lines).*

Discussion. /b/ seems more likely to have tongue constriction by raising and lowering TD, compared to /p/, though not all /b/ tokens were produced with visually apparent TD movement, and some /p/ tokens were still realized with TD movement in ultrasound images. TD constriction direction signaling voicing of bilabial C2s seems to be related to V quality; TD raising is more likely to occur with the high Vs (/i/ and /u/), while TD lowering is more likely to occur with the low V (/ɑ/). This study interprets this finding as the phonetic realization of the voicing of bilabial C2s. TD raises or lowers to potentially create higher aerodynamic pressure in the oral cavity. These TD actions during the acoustic V duration, therefore, enhance the phonetic quality of the preceding V. Specifically, TD raising for /i/ and /u/ narrows the palatal and velar space; TD lowering for /ɑ/ narrows the pharyngeal space during the acoustic V duration before bilabial closure. The lower *f0* of the preceding V of C2 /b/ may be assumed to be the articulatory consequence of TD lowering, which physically discourages the hyoid bone from raising and fronting (Ohala, 1978). However, the models estimated no statistically significant correlation between TD movement distance and direction and *f0* peak values during the acoustic V duration. The acoustic consequence of voicing at the laryngeal action (lower *f0*) may not be directly associated with the articulation of the tongue on bilabial C2s. Regarding TD movement characteristics, voicing coda contrast is apparently accompanied by the spatiotemporal expansion of TD. TD constriction contrasting voicing on bilabial C2s is more evidently characterized by evaluating timing relationships between acoustic and articulatory landmarks. Earlier TD articulation for C2 /b/ from the acoustic V end indicates that AE's phonological voicing can relate to the *gestural aggregation* (Munhall & Löfqvist, 1992) of the laryngeal properties of V with the supralaryngeal properties of C2. This study suggests that the longer duration of V with C2 /b/ may be the articulatory consequence of the speaker's control to create enough room for greater coarticulation of the laryngeal property of V and supralaryngeal properties of bilabial C2 to signal voicing. In conclusion, TD movement should be regarded as an important articulatory correlate indicating the voicing property of bilabial C2s in AE.

References

Ahn, S. (2018). The role of tongue position in laryngeal contrasts: An ultrasound study of English and Brazilian Portuguese. *Journal of Phonetics*, *71*, 451–467.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1).

Coretta, S. (2020). Longer vowel duration correlates with greater tongue root advancement at vowel offset: Acoustic and articulatory data from Italian and Polish. *The Journal of the Acoustical Society of America*, *147*(1), 245–259.

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models—The R Package pbkrtest. *Journal of Statistical Software*, 59(9).

Hothorn, T., Seibold, H., & Zeileis, A. (2023). *partykit*: A Toolkit for Recursive Partytioning (1.2-20) [Computer software].

Levshina, N. (2020). Conditional Inference Trees and Random Forests. In M. Paquot & S. Th. Gries (Eds.), A Practical Handbook of Corpus Linguistics 611–643.

Maddieson, I. (1997). Phonetic Universals. In W. J. Hardcastle & J. Laver (Eds.), The Handbook of Phonetic Science *(1st ed.)*. Blackwells. 619–639.

Munhall, K., & Löfqvist, A. (1992). Gestural aggregation in speech: Laryngeal gestures. *Journal of Phonetics*, *20*(1), 111–126.

Ohala, J. J. (1978). Production of Tone. In V. A. Fromkin (Ed.), Tone: A Linguistic Survey. Elsevier. 5–39.

Svirsky, M. A., Stevens, K. N., Matthies, M. L., Manzella, J., Perkell, J. S., & Wilhelms-Tricarico, R. (1997). Tongue surface displacement during bilabial stops. *The Journal of the Acoustical Society of America*, 102(1), 562–571.

Vazquez-Alvarez, Y., & Hewlett, N. (2007). The 'Trough Effect': An Ultrasound Study. *Phonetica*, 64(2–3), 105–121.

Wrench, A., & Balch-Tomes, J. (2022). Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut. *Sensors*, *22*(3), 1133, 1-27.

Probing the impact of musical training on the temporal malleability of timing in French

Nicole Benker[1], *Miriam Oschkinat*[1], *Philip Hoole*[1], *Simone Falk*[2], *Simone Dalla Bella*[2]

[1]Institute for Phonetics and Speech Processing, Ludwig Maximilian University, Munich, Germany
[2]BRAMS laboratory, University de Montréal, Montréal, Canada

nicole.benker@campus.lmu.de, miriamo@phonetik.uni-muenchen.de

Introduction. Human speech involves the precise interplay of the motor system, the neurological system and sensory input. Thus, a major interest in research on speech production is to study how auditory feedback is integrated into motor planning in subsequent speech. In the current study, we test the malleability of temporal parameters of speech with auditory feedback perturbations by manipulating singleton onsets, vowels and singleton codas in French. Previous research has shown that when speech is temporally stretched or compressed in the auditory feedback, speakers compensate in the opposite direction of the applied perturbation (Floegel, Fuchs, & Kell, 2020; Karlin & Parrell, 2022; Oschkinat & Hoole, 2020) analogously to spectral auditory feedback perturbations (Houde & Jordan, 1998). It has also been found that compensatory responses can persist after perturbation is removed (adaptation), which points towards an update of the underlying motor plan (Oschkinat & Hoole, 2020). Furthermore, it has been shown that stretching of sounds in the signal leads to slowing in following segments (reactive feedback control, Oschkinat & Hoole, 2022). However, compensation and adaptation have not been found consistently across studies. A variety of factors that may influence a compensatory temporal response have been suggested, such as position within the syllable, with complex syllable onsets in German being seemingly less malleable than vowels or complex codas (Oschkinat & Hoole, 2020); crossing of phoneme boundaries, with greater responses when perturbations fall near/across a category boundary (Karlin & Parrell, 2022; Mitsuya, MacDonald, & Munhall, 2014); or the stress pattern of the observed language, with stressed syllables showing greater responses than unstressed syllables (Oschkinat & Hoole, 2022). In addition, individual capacities in non-speech rhythmic behavior and auditory acuity (as for example acquired in musical education) were shown to affect responses, whereby speakers with higher auditory acuity compensated more to ongoing perturbations (online compensation/reactive feedback control) and speakers with higher general motor variability adapted more (Oschkinat, Hoole, Falk, & Dalla Bella, 2022). Except for individual abilities, all of these factors depend on the phonological system of the given language, and have so far been explored for German (Floegel et al., 2020; Oschkinat & Hoole, 2020, 2022) and English (Karlin, Naber, & Parrell, 2021; Karlin & Parrell, 2022; Mitsuya et al., 2014), which are both considered stress-timed languages. Yet, to gain a better understanding of cross-language similarities of temporal representations in speech, it is crucial to investigate languages with different prosodic and rhythmic structure. Accordingly, the language of investigation in the current study is French, which is traditionally regarded as syllable-timed. Additionally, we examine speakers with extensive musical education and compare them with speakers with no (or very little) musical education. In doing so, we test the influence of well-trained feedback-feedforward integration in non-speech tasks (as achieved by musicians with high proficiency on an instrument or in singing) on temporal auditory-motor control in fluent speech.

Methods. Two groups of French speaking participants from the Montréal area (20 musicians and 18 non-musicians, matched in age) completed a real-time temporal auditory feedback adaptation paradigm with two conditions. The paradigm consisted of 95 trials, with four phases (Baseline – no perturbation, Ramp phase – increasing perturbation, Hold phase – maximum perturbation, Aftereffect – no perturbation). In both conditions, participants read the sentence "J'épèle [target word] lundi". In the ONSET condition, the onset /s/ was stretched and the following /u/ delayed and compressed in the target word "soute" (/sut/), and in the CODA condition, the /u/ was stretched and the coda /s/ delayed and compressed in the target word "tousse" (/tus/). If monitoring speech timing in a syllable-timed language is similar to stressed-timed languages, we expect less compensatory shortening to the stretched onset (ONSET condition) than to the stretched vowel (CODA condition), similar to the findings in Oschkinat and Hoole (2020) (H1). This effect could be attributed to a suggested greater articulatory stability and lower malleability of syllable onsets in production (Browman & Goldstein, 1986; Oschkinat & Hoole, 2020). Further, we expect greater compensatory responses to the compressed segments (/u/ ONSET condition, /s/ CODA condition) than to the stretched segments due to effects of reactive feedback control induced by the previous stretched segment (H2). Between the musical groups, we expect stronger compensatory responses in musicians than in non-musicians, under the assumption that musicians engage more efficiently in sensorimotor integration (H3). Durations of the segments /s/ and /u/ were measured and fed into linear mixed-effects models (one per condition) followed by post-hoc tests estimating the contrast of Hold phase productions (with maximum perturbation) and the Baseline (no perturbation), as well as the group differences. Participants for which the perturbation did not work as intended were excluded with an automated MATLAB script.

Results. In the ONSET condition, the group of non-musicians (n=14) showed significant lengthening of the onset /s/ in the Hold phase compared to the Baseline (following the direction of perturbation, b=6.9 ms, p=0.02) while the group of

musicians (n=19) showed no significant effect (b=-3.5 ms, p>0.5). This difference in response between the groups was significant (b=10.3 ms, p = 0.04). Both groups showed significant (compensatory) lengthening of the /u/ (non-musicians: b=7 ms, p=0.005, musicians: b=6.1 ms, p=0.004, non-significant between groups). In the CODA condition, the non-musicians (n=16) significantly lengthened the /u/ (following the perturbation, b=8.8 ms, p=0.034), the musicians (n=18) did not show a significant effect (b=-2.7 ms, p>0.5). Both groups significantly lengthened the coda /s/ (compensatorily, non-musicians: b=29.6 ms, p=0.001, musicians: b=26 ms, p=0.003; Fig. 1). The differences in production between groups were non-significant for both sounds in the CODA condition.

Discussion. The data of the current study do not support our H1: Responses to the stretched /s/ in the ONSET condition were not less pronounced than responses to the stretched /u/ in the CODA condition (Fig. 1, left panels). In fact, unlike in Oschkinat and Hoole (2020), compensatory responses to the vowel in the CODA condition could not be observed at all. H2 is overall supported, given that compensatory responses are much clearer for the compressed segments (right panels), albeit the lengthening responses for the vowel /u/ in the ONSET condition were much weaker than for the /s/ in the CODA condition. Weak responses to the vowels in general could be attributed to their overall shortness in production in the target words, to perturbing within category, and to the characteristics of timing in French. Regarding H3, group differences were observed for the Onset /s/ in the ONSET condition, with musicians indicating a greater tendency towards a compensatory shortening response compared to the non-musicians who lengthened the segment, thereby increasing the perceived auditory mismatch rather than reducing it. Overall, non-musicians seemed to respond more generically by lengthening segments regardless of position within the syllable or perturbation direction, while musicians show a non-significant response rather than following the perturbation. The lack of significant responses in musicians is therefore interpreted as an effort to not generically slow down. These and further results (e.g. for Ramp and Aftereffect phase) will be discussed in more detail at the conference.

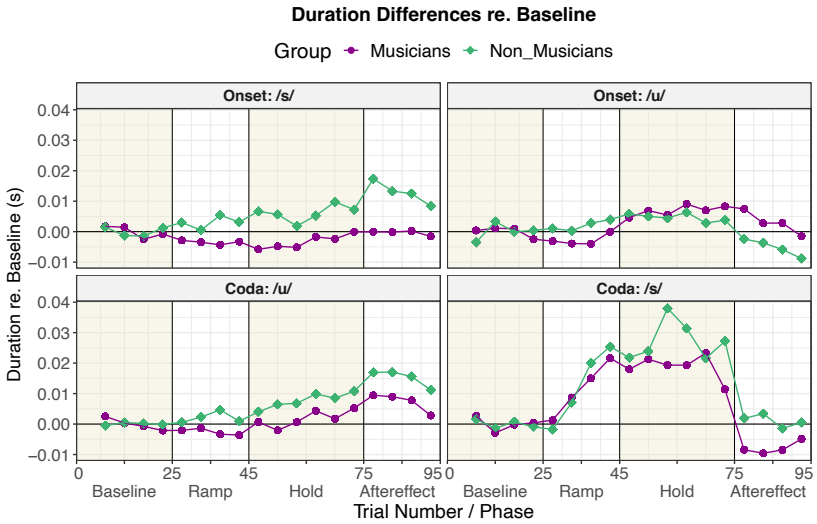


Figure 1. Duration differences (production) relative to the Baseline mean (0) over the course of the experiment binned per 5 trials for musicians and non-musicians. Baseline and Hold phase marked. ONSET condition in the upper panels and CODA condition in the lower panels. Stretched sounds in the left panels, compressed sounds in the right panels. Thus, compensatory responses are indicated by negative values in the left panels and positive values in the right panels.

References

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology, 3*(1), 219-252.

Floegel, M., Fuchs, S., & Kell, C. A. (2020). Differential contributions of the two cerebral hemispheres to temporal and spectral speech feedback control. *Nature Communications, 11:2839*, 1-12.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*(5354), 1213-1216.

Karlin, R., Naber, C., & Parrell, B. (2021). Auditory Feedback Is Used for Adaptation and Compensation in Speech Timing. *Journal of Speech, Language, and Hearing Research*. doi:10.1044/2021_JSLHR-21-00021

Karlin, R., & Parrell, B. (2022). Speakers monitor auditory feedback for temporal alignment and linguistically relevant duration. *The Journal of the Acoustical Society of America, 152*(6), 3142-3154.

Mitsuya, T., MacDonald, E. N., & Munhall, K. G. (2014). Temporal control and compensation for perturbed voicing feedback. *The Journal of the Acoustical Society of America, 135*(5), 2986-2994. doi:10.1121/1.4871359

Oschkinat, M., & Hoole, P. (2020). Compensation to real-time temporal auditory feedback perturbation depends on syllable position. *The Journal of the Acoustical Society of America, 148*(3), 1478-1495. doi:10.1121/10.0001765

Oschkinat, M., & Hoole, P. (2022). Reactive feedback control and adaptation to perturbed speech timing in stressed and unstressed syllables. *Journal of Phonetics, 91*, 101133. doi:https://doi.org/10.1016/j.wocn.2022.101133

Oschkinat, M., Hoole, P., Falk, S., & Dalla Bella, S. (2022). Temporal malleability to auditory feedback perturbation is modulated by rhythmic abilities and auditory acuity. *Frontiers in human neuroscience, 16*. doi:10.3389/fnhum.2022.885074

Acoustic Analysis of Fricatives in Lushootseed

Ted K. Kye[1]

[1]University of Washington
tkye29@uw.edu

Introduction. Although acoustic analysis of fricatives on Salish languages has focused on languages of the Interior Salish branch (Flemming et al. 2008; Gordan et al. 2002; McDowell 2004), there has been little research on fricatives conducted on the Coast Salish branch. The goal of this study is to characterize the acoustic properties of fricatives in Lushootseed, a Coast Salish language in the Puget Sound region with no first language speakers remaining, by analyzing archival recordings dating to the 1950s. Another research goal is to investigate the extent to which the acoustics of fricatives can be analyzed from archival recordings that have an upper frequency cutoff near 5~6kHz (which pose challenges to the analysis of /s/). The recordings used for this study come from the Metcalf collection, which is part of the Burke Museum's Special Collections. Several of these recordings were digitized at 44.1kHz with 16-bit depth. From this collection, recordings of connected speech from two elders were examined: Annie Jack (AJ) and Martha Lamont (ML).

Methods. The recordings were analyzed and annotated through Praat. Fricatives /s ʃ ɬ xʷ χ χʷ/ ‹s š ɬ xʷ x̌ x̌ʷ› were examined. Following Shadle (2012, 2023), spectral moments (Center of Gravity (CoG), kurtosis, skew, and variance) were measured from a time-averaged spectrum, where Discrete Fourier Transforms (DFTs) were extracted from six time points and averaged through a matrix of intensity and sampling frequencies. The length of each window was 15ms. Fricatives that were less than 56ms were omitted to avoid potential overlap across each window. The script that was used for time-averaging is from DiCanio (2021). Intensity was also measured given that fricatives may be expected to differ due to differences in airflow resulting from differences in the area of the constriction. Intensity (in dB) was measured at 5 time points along the fricative duration: 10%, 30%, 50%, 70%, 90%. The data was fit into a linear mixed effects model with spectral moments and intensity as dependent variables, fricatives as fixed effects (using backward difference coding), and speakers as random effects.

Results. The summary statistics in Table 1 summarizes the means and standard deviations of spectral moments for each fricative in Lushootseed.

Table 1: *Means and standard deviations of spectral moments (in Hz) for each Lushootseed fricative.*

Consonant	***N***	***CoG***	***SD***	***Kurtosis***	***SD***	***Skew***	***SD***	***Variance***	***SD***
[s]	253	2766	792.53	20.49	23.32	2.64	1.4	2110	744.11
[ʃ] <š>	47	2262	454.3	123.83	170.37	6.97	4.98	1264	508.82
[ɬ]	135	1376	553.68	77.83	74.24	6.35	2.67	1622	828.84
[xʷ]	208	1027	391.41	448.11	535.00	15.66	8.68	1117	814.46
[χ] <x̌>	44	1294	210.67	256.81	231.33	11.00	4.99	832	387.6
[χʷ] <x̌ʷ>	52	826	165.99	672.65	793.88	18.46	9.19	783	584.7

There was a frequency cutoff near 5~6kHz in the recordings. Studies have shown that /s/ tends to reach its peak amplitude near 8kHz (Shadle 1985; Koenig et al. 2013). Although absolute measures of spectral moments for /s/ could not be measured due to this frequency cutoff, relative differences (based on the first 5~6kHz) across fricative places of articulation were obtained (illustrated in Figure 1a). The relative CoG for /s/ was significantly greater than /ʃ ɬ/, /ʃ ɬ/ greater than /xʷ χ/, and /xʷ χ/ greater than /χʷ/ (i.e., /s/ > /ʃ ɬ/ > /xʷ χ/ > /χʷ/). There were also across-speaker differences, where the CoG for /ʃ/ was significantly greater than /ɬ/ for the speaker ML but not AJ. Moreover, the CoG for /xʷ/ was the same as /χ/ for AJ but the same as /χʷ/ for ML. However, /xʷ/ was differentiated from /χ/ based on the variance from both speakers, where the variance for /xʷ/ was greater (by 285Hz) than /χ/. Moreover, the labialized dorsal fricatives /xʷ χʷ/ was reliably differentiated from /χ/ based on the kurtosis, which was significantly greater than /χ/. The fricative /ɬ/ can also be differentiated from /ʃ/ based on the variance, where the variance for /ɬ/ was significantly greater (by 294Hz) than /ʃ/. Skew differentiated backness vs. frontness, where the skew for /s/ was smallest, largest for the dorsal fricatives /xʷ χ χʷ/, and in-between for /ʃ ɬ/ (i.e., /xʷ χ χʷ/ > /ʃ ɬ/ > /s/). Unexpectedly, given that /s/ is a sibilant, intensity for /s/ was the lowest, whereas the intensity of /χ χʷ/ was highest (the levels of intensity were /χ χʷ/ > /xʷ/ > /ʃ/ > /ɬ/ > /s/). Figure 1b illustrates the intensity (in dB) for each fricative at the five time points.

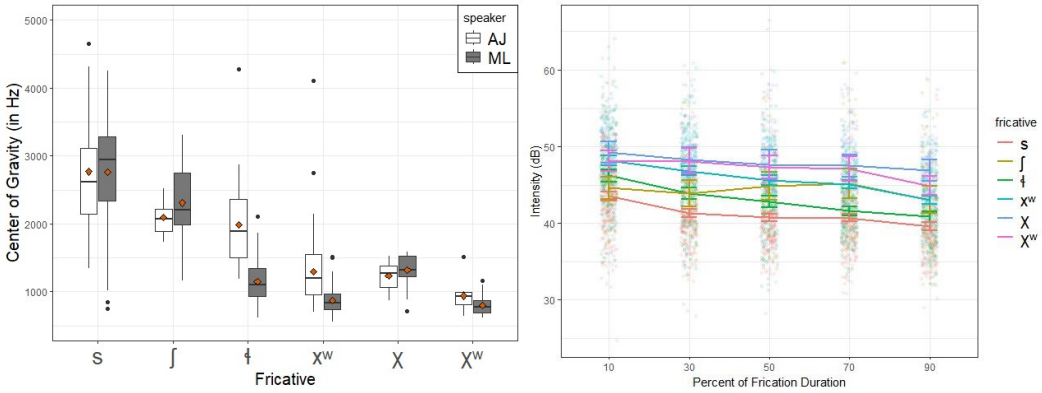


Figure 1: (a) *Boxplot for time-averaged CoG for each fricative in Lushootseed across the two speakers, and (b) intensity of each fricative at five time points.*

Discussion.

Spectral moment measurements reliably differentiated fricative contrasts in Lushootseed. Where one measure didn't show a contrast, the other did. For example, although /xʷ/ and /χ/ did not differ in CoG, they differed in variance, where the variance for /xʷ/ was greater than /χ/. The labialized dorsal fricatives showed higher kurtosis than the plain uvular fricative /χ/, which suggests that the acoustic coupling of lip rounding yields a more narrow spectral peak. Skew was found to be a good predictor of the frontness/backness dimensions of the articulators, where skew for /s/ was the smallest, dorsal fricatives /xʷ χ χʷ/ largest, and /ʃ ɬ/ in-between.

The current findings appear to provide evidence for a difference in the realization of /ɬ/ when compared with other Salish languages (i.e., Montana Salish), where /ɬ/ tends to be closer to /ʃ/ (Gordon et al. 2002). In contrast, the lateral fricative /ɬ/ has a lower CoG than /ʃ/ for the speaker ML. It is possible that the articulatory release for the lateral fricative was made more posteriorly along the sides of the palate for this speaker. The more posterior the constriction, the lower the center of gravity (Gordon et al. 2002). Evidence of retraction in lateral obstruents (/ɬ/ and /tɬ'/) has been observed from ultrasound imaging of Montana Salish (McDowell 2004). However, retraction may not account for the speaker ML because the Montana Salish retraction did not corroborate with the acoustic findings of /ɬ/ in Montana Salish, where the CoG for /ɬ/ was (on average) greater than /ʃ/ (Gordon et al. 2002). Another possibility for the low CoG in the current data is that it is due to a difference in the length of the buccal cavity during the release. The lower CoG is not observed for the speaker AJ, where the CoG for /ɬ/ was (as expected) approximately the same as /ʃ/. This suggests that the contrast may be due to cross-speaker differences in the production of /ɬ/ rather than a genuine cross-linguistic difference.

The intensity of dorsal fricatives was greater than sibilants. It should be noted, however, that dorsal fricatives are expected to have lower intensity than sibilants. The area of the constriction for dorsal fricatives is greater than articulations made with the tongue tip or tongue blade because of the greater mass of the tongue body. The airflow through the constriction is lower because of the greater area of constriction between the rounded tongue dorsum and the rounded soft palate, which would subsequently lead to a drop in intensity (Kent & Moll 1972). However, the upper frequency cutoff from the microphone signal made the peaks above 5~6kHz lost for sibilants. This suggests that intensity may not be a reliable measure for characterizing the fricative contrasts from archival recordings like these.

References

DiCanio, Christian (2021). Spectral moments of fricative spectra script in Praat. Retrieved from https://www.acsu.buffalo.edu/~cdicanio/scripts/Time_averaging_for_fricatives_4.0.praat.

Flemming, Edward, Peter Ladefoged, & Sarah Thomason (2008). Phonetic structures of Montana Salish. *Journal of Phonetics*, *36*(3), 465-491.

Gordon, Matthew, Paul Barthmaier, & Kathy Sands (2002). A cross-linguistic acoustic study of voiceless fricatives. In *Journal of the International Phonetic Association, 32*(2), 141-174.

Kent, Raymond & K. Moll (1972). "Cineflourographic analyses of selected lingual consonants." In *Journal of Speech and Hearing Research, 15*, 453-473.

Koenig, Laura L., Christine H. Shadle, Jonathan L. Preston, & Christine R. Mooshammer (2013). Toward improved spectral measures of /s/: Results from adolescents. In *Journal of Speech, Language, and Hearing Research, 56*, 1175-1189.

McDowell, Ramona E. (2004). *Retraction in Montana Salish lateral consonants* (Doctoral dissertation, University of British Columbia).

Shadle, Christine H. (1985). The Acoustics of Fricative Consonants. (Doctoral Dissertation, Cambridge Massachusetts).

Shadle, Christine H. (2012). Acoustics and aerodynamics of fricatives. In A. Cohn, C. Fougeron, and M Huffman (eds.) *Handbook of Laboratory Phonology*, pp. 511-526. Oxford University Press, Oxford, UK.

Shadle, Christine. H. (2023). Alternatives to moments for characterizing fricatives: Reconsidering Forrest et al. (1988). In *The Journal of the Acoustical Society of America*, *153*(2), 1412-1426.

Dentition and Articulations of Mandarin Rhotic /ɹ/ and Retroflex /ʂ̩/: A Preliminary Study

Yung-hsiang Shawn Chang[1], Jeyang Jau[2]

[1]Department of English, National Taipei University of Technology
[2] O2Win Dental Clinic
shawnchang@mail.ntut.edu.tw, jeyang@yayi.tw

Introduction. English rhotic /ɹ/ is recognized for its articulatory variability, with varying tongue postures generally categorized as retroflex (tongue tip pointing up) or bunched (tongue tip pointing down) (e.g., Delattre & Freeman, 1968; Boyce & Espy-Wilson, 1997). Previous research has reported the influences of phonetic contexts (e.g., Guenther et al., 1999), dialects (e.g., Delattre & Freeman, 1968), and palatal shape/size (Bakst & Lin, 2014) on the articulation of English /ɹ/. Similar to palatal morphology, dental occlusion has also been found to impact articulation. Available studies have focused the association between malocclusion and speech distortions (e.g., Amr-Rey et al., 2022; Assaf et al., 2021), particularly in individuals with Angle Class II malocclusion (overbite) and Class III (underbite). While the adverse impact of malocclusion on speech articulation is evident, the connection between occlusion types and specific articulatory choices in sounds characterized by variable articulatory gestures remains unclear. This study aims to address this gap by conducting an exploratory investigation with Mandarin rhotic /ɹ/ and retroflex /ʂ̩/, which both have been found to feature tongue posture variations similar to English /ɹ/ (e.g., Chang, 2018; Chen & Mok, 2021).

Methods. Twenty-two Taiwan Mandarin speakers, self-reported to have normal dental occlusion, were recruited from the first author's institution. Additionally, forty-three Taiwan Mandarin-speaking participants were recruited through the second author's clinic, with 14 diagnosed as Class I malocclusion, 13 as Class II, and 16 as Class III. In light of prior research (e.g., Leavy et al., 2016) that indicated no noticeable speech issues in individuals with Angle's Class I malocclusion, this study focused exclusively on the data of participants diagnosed with Class II and Class III malocclusion. The stimuli consisted of syllabic rhotic /ɻ̩4/ and syllabic fricative /ʂ̩2/ (representing "two" and "ten" in Mandarin, respectively), repeated 12 times in a number-reading task. Data collection took place in a quiet room, using a CHISON ECO1 portable ultrasound machine. Given potential variations in the pronunciation of /ɻ̩/ and /ʂ̩/ by Taiwan Mandarin speakers, such as substituting them with /ə/ and /ʂ/ respectively (e.g., Lin, 2007), all participants' data were screened by two phonetically trained research assistants based on auditory impression. The screening results revealed that, with the exception of two speakers with normal occlusion and one speaker with Cass III malocclusion, all other participants produced intelligible /ɻ̩/ and /ʂ̩/ tokens. These three participants' data were excluded from subsequent analyses. Ultrasound images were extracted from the midpoint of the syllabic /ɻ̩/ and /ʂ̩/ productions and then categorized into bunched and retroflex shapes following Chen & Mok (2021).

Results. Similar to the findings of Chen & Mok (2021) and Chang (2018), we observed various shapes for the bunched posture—front bunched, back bunched and back bunched with a dip—and for the retroflex posture—tip-up retroflex and tip-up-and-backward retroflex for Mandarin rhotic /ɹ/ and fricative /ʂ/ (see Figure 1). However, for the sake of comparison with earlier articulatory studies on rhotic and retroflex sounds, we simplified the categorization of tongue postures to either as bunched or retroflex. It should be noted that all participants nearly exclusively used either the bunched or retroflex tongue posture across repetitions of the same token. As seen in Table 1, the bunched posture is predominant for Mandarin /ʂ/ across all three groups of participants. The retroflex posture emerges as the preferred articulatory gesture for Mandarin /ɹ/ for individuals with normal occlusion, but it is used less among individuals with malocclusion.

Table 1: *Tongue shapes for Mandarin rhotic /ɹ/ and fricative /ʂ̩/ across occlusion types.*

	normal	Class II malocclusion	Class III malocclusion
/ɹ/	Retroflex: 60% Bunched: 40%	Retroflex: 46.2% Bunched: 53.8%	Retroflex: 26.7% Bunched: 73.3%
/ʂ̩/	Retroflex: 20% Bunched: 80%	Retroflex: 30.8% Bunched: 69.2%	Retroflex: 0% Bunched: 100%

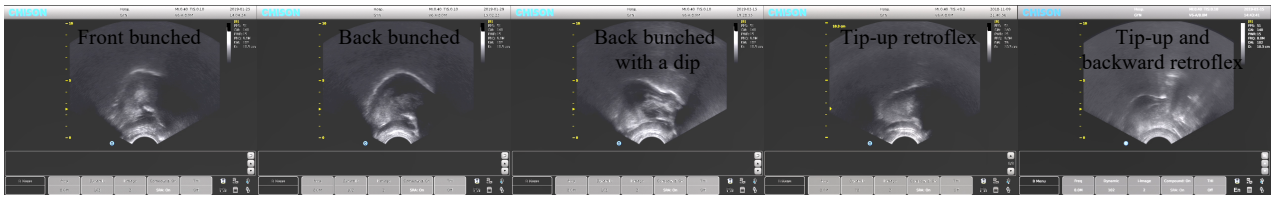


Figure 1: *Tongue shapes for Mandarin* rhotic /ɹ/ and fricative /ʂ/

Discussion. The current study undertook a preliminary investigation into the relationship between the types of dental occlusion and choices of articulatory gestures in producing articulatorily variable Mandarin rhotic /ɹ/ and fricative /ʂ/. Consistent with the findings of Farronato et al. (2012), our results show that individuals with Class II malocclusion, compared to Class III malocclusion, exhibit retroflex vs. bunched variations more comparable to those with normal occlusion. Importantly, we observed that individuals with Class III malocclusion used the retroflex gesture much less than the other two participant groups. It is speculated that the retroflex tongue gesture, which involves the tongue tip approximating the anterior part of the hard palate, is challenging to execute with a protruding lower jaw in Class III individuals. Additionally, the fricative /ʂ/ requires a sufficiently large cavity anterior to the lingual constriction for sibilancy, potentially explaining why speakers with a relatively small maxilla due to Class III occlusion disfavor a retroflex tongue posture.

Although articulatory variants of English rhotic /ɹ/ are largely indistinguishable acoustically and perceptually (e.g., Twist et al., 2007; Westbury et al., 1998), the retroflex vs. bunched variants of Mandarin retroflexes have been shown to be perceptually and acoustically distinct (Ou & Chen, 2014). Whether these variants are distinguishable acoustically and perceptually across different types of occlusion is still not known and will be explored in the next step of our study.

References

Amr-Rey, O., Sánchez-Delgado, P., Salvador-Palmer, R., Cibrián, R., & Paredes-Gallardo, V. (2022). Association between malocclusion and articulation of phonemes in early childhood. *The Angle Orthodontist*, *92*(4), 505-511.

Assaf, D. D. C., Knorst, J. K., Busanello-Stella, A. R., Ferrazzo, V. A., Berwig, L. C., Ardenghi, T. M., & Marquezan, M. (2021). Association between malocclusion, tongue position and speech distortion in mixed-dentition schoolchildren: an epidemiological study. *Journal of Applied Oral Science*, *29*.

Bakst, S., & Lin, S. (2015). An ultrasound investigation into articulatory variation into American /s/ and /r/. In: Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015).

Boyce, S., & Espy-Wilson, C. Y. (1997). Coarticulatory stability in American English /r/. *The Journal of the Acoustical Society of America*, *101*(6), 3741-3753.

Chang, Y.-H. S. (2018). Articulatory and acoustic investigations into gestures of Mandarin retroflex fricatives. Poster presented at the 176th Meeting of the Acoustical Society of America, Victoria, Canada.

Chen, S., & Mok, P. P. K. (2021). Articulatory and Acoustic Features of Mandarin /ɹ/: A Preliminary Study. In *Proceedings of 12th International Symposium on Chinese Spoken Language Processing (ISCSLP).*

Delattre, P. C., & Freeman, D. C. (1968). A dialect study of American rs by x-ray motion picture. *Linguistics*, *44*, 29-68.

Farronato, G., Giannini, L., Riva, R., Galbiati, G., & Maspero, C. (2012). Correlations between malocclusions and dyslalias. *European Journal of* Paediatric Dentistry, *13*(1), 13-18.

Ou, S. C., & Guo, Z. C. (2014). Mandarin retroflex sounds perceived by non-native speakers. *Journal of Language and Literature Studies*, *26*, 39-76.

Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., & Perkell, J. S. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/production. *Journal of the Acoustical Society of America*, *105*, 2854–2865.

Leavy, K. M., Cisneros, G. J., & LeBlanc, E. M. (2016). Malocclusion and its relationship to speech sound production: Redefining the effect of malocclusal traits on sound production. *American Journal of Orthodontics and Dentofacial Orthopedics*, *150*(1), 116-123.

Lin, Y. H. (2007). The *Sounds of Chinese*. Cambridge: Cambridge University Press.

Twist, A., Baker, A., Mielke, J., & Archangeli, D. (2007). Are 'covert' /r/ allophones really indistinguishable?. *University of Pennsylvania Working Papers in Linguistics, 13*(2), 207-216.

Westbury, J. R., Hashi, M., & Lindstrom, M. J. (1998). Differences among speakers in lingual articulation for American English/ɹ. *Speech Communication*, *26*(3), 203-226.

Prosodic variation of kibushi dialects (Mayotte, France) via the reading of *The Little Prince*: a pilot study

Ahamada Kassime[1], Fabrice Hirsch[1], Miki Mori[1,2]

[1]UMR 5267 Praxiling – University Paul-Valéry Montpellier 3 & CNRS
[2]CUFR Mayotte

ahamadakassime5@gmail.com, fabrice.hirsch@univ-montp3.fr, miki.mori@univ-mayotte.fr

Introduction. Mayotte, situated in the Mozambique Channel between Madagascar and the Comoros Archipelago, is a multilingual French island. Its sociolinguistic landscape is intricately tied to its geographical location, characterized by two principal local languages: Shimaore, a Sabaki Bantu language spoken by the majority, and Kibushi, an Austronesian language used by only 15% of the Maore population (Insee, 2022). Furthermore, having transitioned from a French colony in the second half of the 18th century to a French Department in 2011, Mayotte recognizes French as its sole official language for administration and education.

Concerning Kibushi, this language exhibits linguistic affinity with a Malagasy dialect from the northwest of Madagascar, the Sakalava Dialect (Bare-Thomas, 1976). Within Mayotte, there are two Kibushi dialects spoken across 21 villages: the Kisakalava dialect, present in 18 villages, and the Kiantalautsi Dialect, spoken in only 3 villages (Jamet, 2016). Research on Kibushi has also revealed intra-dialectal variations (Gueunier, 2004; Jamet, 2016; Laroussi, 2010), indicating that each village is linguistically autonomous, employing distinct prosodic parameters such as accent, intonation, rhythm, and lexicon to define these variations.

A recent study on Kibushi stress variation revealed diatopic differences, within three prosodic parameters— duration, pitch, and intensity — used to compare syllable stress in words across villages and dialects (Kassime *et al.*, 2023). However, no research has been carried out to determine the prosody of Kibushi in utterances longer than the word. With this in mind, the aim of this study is to describe the rhythmic and melodic organization of the different varieties of kibushi, focusing on the statements. Our hypothesis is that Kibushi prosody differs in terms of pitch and syllable length according to the dialect spoken and the village in which it is used.

Methods. Data were acquired in three villages: Acoua and Chiconi (Kisakalava dialects), and Ouangani (Kiantalautsi dialect). A total of 26 participants were recorded, with 11 from Acoua (6 men and 5 women), 7 from Chiconi (1 man and 6 women), and 7 from Ouangani (6 men and 1 woman). The participants, with an average age of 16, had to read passages from *The Little Prince* by Antoine de Saint-Exupéry in Kibushi, totaling 3,464 words. Recordings were transcribed on 3 tiers in Praat (Boersma and Weenink, 2019): an orthographic tier, a syllabic tier and a phonemic tier. Prosodic variation results were obtained through analyses and graphics generated using Prosogram (Mertens, 2022), a tool designed for the analysis of pitch variations in speech. An extension of this tool, called Polytonia (Mertens, 2014), was used to obtain the highest and lowest points of F0. Statistical analyses were carried out using R software. Additionally, to account for differences between male and female voices, fundamental frequency (f0) was converted from Hertz to semitones for each participant, using each participant's f0 averages (12*(log((zz)/100)) / log(2)).

The following measures have been taken:

- The average fundamental frequency (F0);
- The average minimum fundamental frequency (F0_min);
- The average maximum fundamental frequency (F0_max);
- The highest point of the fundamental frequency;
- The lowest point of the fundamental frequency;
- Syllable duration (in ms).

Results.

Pitch variations in speech

Firstly, the results reveal no significant difference between f0_min and f0_max for each participant, village, and dialect. However, when comparing pitch variations by participant and dialect, a subtle distinction emerges. Three participants (F12, H8 from Acoua, and F2 from Chiconi) stand out from the others, exhibiting lower high and low pitch. Furthermore, the Kiantalautsi dialect (Ouangani=f0_min: mean -0.04771 ST; f0_max: mean -0.0556 ST) displays a slight disparity

from the two Kisakalava dialects (Chiconi=f0_min: mean -0.1300 ST; f0_max: mean -0.1512 / Acoua=f0_min: mean -0.1178 ST; Acoua=f0_max: mean -0.1417 ST) with higher pitch.

High pitch in statements

Regarding proeminence in statements, the subject-predicate order, predominantly employed by all participants, was chosen for analysis. The results indicate that the high pitch (H) is generally situated on the predicate across all villages. Specifically, Chiconi exhibited a realization of 77% (n=27) H pitch on the predicate, Acoua demonstrated 79% (n=33), while Ouangani showed a slightly lower percentage at 63% (n=35). Notably, these high pitches are predominantly located on penultimate and ultimate syllables.

Furthermore, a subtle difference was observed in Ouangani, where 51% (n=47) of H pitch occurred on ultimate syllables, in contrast to 34% for both Chiconi (n=32) and Acoua (n=41).

Syllable stress duration

In terms of syllable length, stressed syllables consistently exhibit longer durations than unstressed ones across all three villages. The median duration of unstressed syllables does not significantly differ between the two dialects: 183 ms for Chiconi (min.=21 ms; 1st Qu.=141 ms; median=183 ms; mean=217 ms; 3rd Qu.=254 ms; max.=876 ms), 180 ms for Acoua (min.=18 ms; 1st Qu.=138 ms; median=180 ms; mean=217 ms; 3rd Qu.=248 ms; max.=850 ms) , and 173 ms for Ouangani (min.=21 ms; 1st Qu.=136 ms; median=173 ms; mean=185 ms; 3rd Qu.=224 ms; max.=680 ms).

However, in stressed syllables, Chiconi stands out with a notably longer duration (median=324 ms) (min.=45 ms; 1st Qu.=225 ms; median=324 ms; mean=353 ms; 3rd Q.=454 ms; max.=1006 ms). In comparison, Acoua and Ouangani have slightly shorter median durations for stressed syllables, with Acoua at 274 ms (min.=14 ms; 1st Qu.=198 ms; Median=274 ms; Mean=316 ms; 3rd=Qu. 386 ms; Max. 1324 ms) and Ouangani at 256 ms (min.=38 ms; 1st Qu.=180 ms; median=256 ms; mean=288 ms; 3rd Qu.=349 ms; max.=1052 ms).

Discussion. Results obtained from the two prosodic parameters, pitch and duration, suggest that Kibushi variation depends on villages and dialects.

In a reading context, pitch exhibits slight variations. Specifically, the Kiantalautsi dialect (Ouangani) differs slightly from the two Kisakalava dialects (Acoua and Chiconi) in terms of high (H) vs low (L) pitch and the location of H pitch in declarative sentences. Regarding syllable stress duration, readers from Chiconi produced longer durations on stressed syllables than those from the other villages. This phenomenon may contribute to the perception by some Kibushi speakers that individuals from Chiconi speak slowly (Kassime, 2021).

To further enhance our understanding of prosodic variation in Kibushi, future research is essential. Expanding the dataset to include a wider range of declaratives and wh-questions will contribute to a more comprehensive exploration of these findings.

References

Bare-Thomas D. (1976). Le dialecte sakalava du nord-ouest de Madagascar : phonologie, grammaire, lexique. Linguistic dissertation, Université de Paris V.

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer (version 6.1.01). http://www.praat.org

Gueunier, N. (2004). Le dialecte malgache de Mayotte (Comores) : une discussion dialectologique et sociolinguistique. *Faits de Langues*, 23-24, 397-420.

Insee .(2022). Mayotte, un territoire riche de ses langues et de ses traditions. Insee Analyses Mayotte, 33. https://www.insee.fr/fr/statistiques/6467148

Jamet, R. (2016). Dictionnaire kibushi-français, Dzoumogné, Association SHIME.

Kassime, A. (2021). Représentations et attitudes autour des variables du kibushi kisakalava et leur place dans un territoire plurilingue (Mayotte). Université Rennes 2.

Kassime, A., Hirsch, F., & Mori, M. (2023). Vowel space and acoustic characteristics of stressed syllables in two kibushi dialects spoken in Mayotte, France: a pilot study. *20th International Congress of Phonetic Sciences (ICPhS 2023)*, Aug 2023, Prague, Czech Republic. pp.3321-3324. (hal-04198178).

Laroussi, F. (2010). Langues, identités et insularité. Regards sur Mayotte. Université de Rouen et du Havre.

Mertens P. (2014). Polytonia: a system for the automatic transcription of tonal aspects in speech corpora. *Journal of Speech Sciences,* 4(2), 17-57.

Mertens, P. (2022). The Prosogram model for pitch stylization and its applications in intonation transcription. in Barnes, J.A. and Shattuck-Hufnagel, S. (eds) (2022) Prosodic Theory and Practice. Cambridge, MA: MIT Press 259-286. ISBN 978-0-262-54317-0.

EEG dynamics and source identification during air volume reduction induced by a long utterance

Said-Iraj Hashemi[1,2], Guy Cheron[1], Didier Demolin[2], Ana Maria Cebolla Alvarez[1]

[1]Laboratory of Neurophysiology and Movement Biomechanics (LNMB). Faculty of Motor Skills Sciences | Motricity Sciences Research Center. Interfaculty institutes | UNI, ULB Neuroscience, Institute Université Libre de Bruxelles, Brussels, Belgium
[2]Phonetics and Phonology Laboratory LPP, CNRS-UMR 7018, Sorbonne Nouvelle, Paris, France
Said-iraj.hashemi@ulb.be

Introduction

During phonation, several respiratory muscles contribute to maintaining a quasi-constant subglottic pressure (Ps) beneath the vocal folds in short utterances. However, for longer utterances, it becomes imperative to involve one or more expiratory muscles to maintain the level of Ps required for phonation. Ladefoged [1] highlighted the different actions of the respiratory muscles responsible for maintaining Ps during a long utterance. After a deep inspiration, the lung volume is big, then the inspiratory muscles relax which triggers the elastic recoil of the inspiratory muscles and the lung volume decreases. When the lung volume becomes low, the Ps required for speech is maintained by the work of the expiratory muscles. Toward the end of a long utterance, when the volume of air in the lungs is very low, many expiratory muscles may be needed to maintain Ps at a sufficient level. We hypothesize that specific brain oscillatory activity underlies these phases of muscular sequential activations during long utterances. In this study, we investigated the electroencephalogram cortical patterns of power spectrum perturbations and the related brain generators of the alpha brain rhythm, associated with the regulation of Ps and underlying muscular actions during a long utterance.

Methods

We recruited 19 French-speaking students (10 men, 9 women) from the l'Université Libre de Bruxelles in good health. The protocol involved a simple action: taking a deep breath and pronouncing the syllable [pa] as long as possible without taking another breath. Each participant completed 20 trials of the task. Participant kept their eyes closed during each trial. This work was approved by the Ethical Committee of Hôpital de Brugmann, Belgium.

During each trial, we measured intraoral pressure and oral airflow rate. Electromyographic activity (EMG) from respiratory muscle activity was recorded with bipolar Delsys EMG TrignoTM surface electrodes placed on the scalene, external, and internal intercostal, external oblique, and rectus abdominis muscles.
EEG signals during [pa] sequences were recorded using the ANT system (sampling rate 2048 Hz), (Netherlands) Neuro with a 64-channel. In post-recording, all electrodes are re-referenced in REST (Reference Electrode Standardization Technique). All recording systems were synchronized with each other using an external signal generator sending triggers at the beginning and end of each recording.
All the processing was done using MATLAB and the EEGlab [2] toolbox. Brain generator identification was performed using a distributed method (the standardized weighted Low Resolution Electromagnetic Tomography, swLORETA [3]) in the ASA software (ANT neuro).

Results

Aerodynamic and muscle activity

Careful visual inspection of the respiratory EMGs involved in PS regulation allowed us to identify the three distinct phases previously described by Titze [4] and later identified by us as four (Figure 1). The first phase (P1) corresponds to the elastic recoil of the external intercostal muscle, phase 2 (P2) represents the delay between the end of the elastic recoil and the onset of external oblique activity (which likely corresponds to the internal intercostal activity), phase 3 (P3) is characterized by external oblique activity, and finally, phase 4 (P4) corresponds to activity in the rectus abdominis.

EEG and source analysis
Preliminary analysis of EEG signals included filtering (0.5 Hz high pass, 200 Hz low pass), and the rejection of artefactual components (ocular movement components and muscle activity) detected by ICA analysis (eeglab). Data were epoched

to the events (at 0 s). labeling each one of the four phases at 0 s. We focused our analysis on the brain source localization of the alpha rhythm (10.5 to 13 Hz). Following non-parametric permutations and Holm post hoc tests, the model revealed that the alpha rhythm generators were localized in classical pre-motor and motor areas during the four phases. Concretely, we found that the first phase corresponding to the elastic recoil of the external intercostal muscle was characterized by the activation of the right primary motor and left primary somatosensory cortex, the second phase, related to the action of the internal intercostal muscle, showed activation in the left premotor cortex, supplementary motor cortex, and the left agranular retrolimbic area. Similarly, phase 3, linked to the action of the external oblique muscle, exhibited activity in the left and right premotor and supplementary motor cortex, and left dorsolateral prefrontal cortex. Phase 4, associated with the contraction of the rectus abdominis muscle, was characterized by left opercular Broca's area involvement.

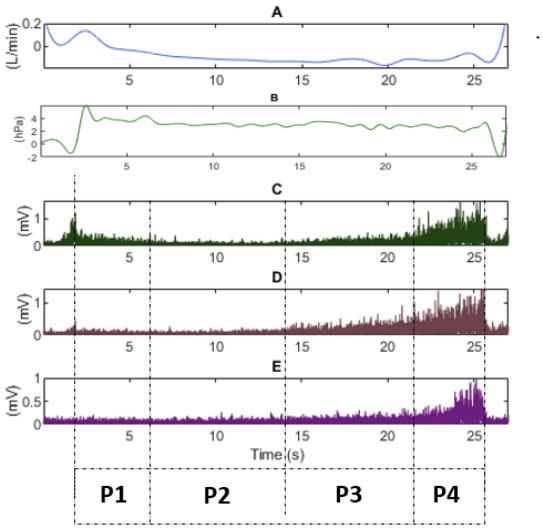


Figure 1: Aerodynamic dynamics and muscle activity. A:air flow; B:intra-oral pressure; C:external intercostal; D:external oblique; E:rectus

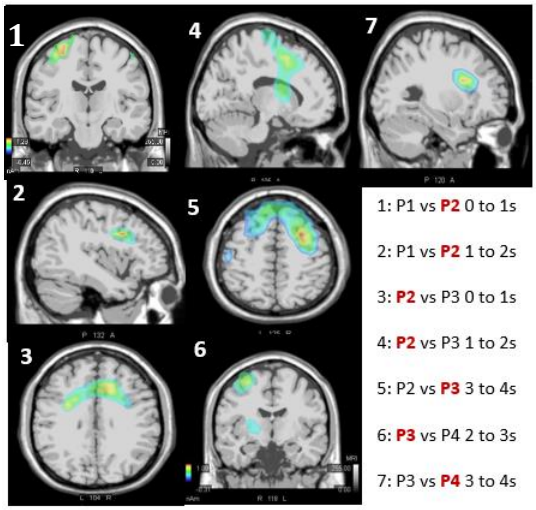


Figure 2 : Sources identification; P: Phase to which the identified source is associated

In Figure 2, in squares, we can observe **1**: right primary motor and left primary somatosensory cortex, **2**: Right front eye fields, **3**: left premotor and supplementary motor cortex, **4**: left premotor and supplementary motor cortex, left caudate and right primary motor (not shown), **5**: Right premotor and supplementary motor cortex, left dorsolateral prefrontal cortex, **6**: Right premotor and supplementary motor cortex and right globus pallidus, **7**: Left Broca-opercular

Discussion

The present results revealed classical motor and premotor areas involvement during the four phases which were defined by the specific respiratory muscle activations during phonation. This suggests continuous cortical control or monitoring during long utterances. Interestingly, the left dorsolateral prefrontal cortex was involved during phase 3. As this region is recognized to underlie cognitive processes as those related to decision-making, it could be hypothesized that this phase has a motoric volitional character that may be linked to the subsequent involvement of Broca area found in phase 4 for succeeding phonation during the last seconds characterized by the collective muscular effort of all the expiratory muscles, as well as the prolonged contraction time of the jaw and facial muscles required to produce the syllable [pa]. According to the study by Ferpozzi et al. (2018) [5], Broca's area is a functional gate authorizing the phonetic translation to be executed by the motor areas**.**

References

1: Ladefoged P. Three areas of experimental phonetics: stress and respriatory activity, the nature of vowel quality, units in the perception and production of speech. --. London: Oxford U.P.; 1967.

2: Makeig S, Debener S, Onton J, Delorme A. Mining event-related brain dynamics. Trends in Cognitive Sciences 2004;8:204–10. https://doi.org/10.1016/j.tics.2004.03.008.

3: Cebolla AM, Palmero-Soler E, Dan B, Cheron G. Frontal phasic and oscillatory generators of the N30 somatosensory evoked potential. Neuroimage. 2011 Jan 15;54(2):1297-306. doi: 10.1016/j.neuroimage.2010.08.060. Epub 2010 Sep 8. PMID: 20813188.

4. Titze, I. Principles of Voice Production. Prentice Hall. 1994.

5: Ferpozzi V, Fornia L, Montagna M, Siodambro C, Castellano A, Borroni P, Riva M, Rossi M, Pessina F, Bello L, Cerri G. Broca's Area as a Pre-articulatory Phonetic Encoder: Gating the Motor Program. Front Hum Neurosci. 2018 Feb 22; 12:64. doi: 10.3389/fnhum.2018.00064. PMID: 29520225; PMCID: PMC5826965.

Laterals in simplex vs. complex syllable codas: a comparison of four languages

Anisia Popescu[1], Ioana Chitoran[2]

[1]LISN, Université Paris Saclay
[2]Clillac-Arp, Université Paris Cité
anisia.popescu@universite-paris-saclay.fr

Introduction. The present paper investigates coda coordination patterns as a function of /l/ darkness in Russian, English, Romanian and Georgian. Since first described within the framework of Articulatory Phonology (Browman & Goldstein, 1988) syllable level coordination patterns have been the focus of many studies. More attention has been given to cross-linguistic differences in gestural coordination in onsets, which are hypothesized to have a global coordination pattern (i.e., consonant gestures are synchronically timed with the vowel gesture). To our knowledge, far fewer studies have looked at between language differences of coordination patterns in codas, which are hypothesized to have a local organization (i.e, the vowel nucleus and coda consonants are sequentially timed). Coda clusters involving the lateral consonant in American English (and the rhotic trill in Romanian) have been found to exhibit a diverging global coordination pattern in coda position (Marin & Pouplier, 2010; 2014). Marin and Pouplier (2014) suggest that the articulatory characteristics of the liquids trigger the global coordination patterns found in coda position. The coda lateral in American English is a dark /l/, and is produced with a double lingual gesture: a vocalic tongue dorsum (TD) retraction that precedes a consonantal tongue tip (TT) raising. Similar to the dark /l/, the rhotic trill is also produced with a double gesture (Proctor, 2009): a vocalic tongue root (TR) gesture that precedes and acts as an anchor for the consonantal TT trilling gesture. Thus, the common gestural specifications of the American English /l/ and the Romanian /r/, the cases where coda global organization was found, both share the presence of a double lingual gesture and an earlier occurring vocalic gesture. We therefore suggest that global organization in coda position occurs because of the existence of an earlier vocalic gesture that triggers gestural competition between the vowel nucleus and the vocalic gesture of the liquid.

To test this hypothesis, the present paper compares coda coordination patterns in four languages that differ in their gestural synergies of their coda lateral consonant: Russian (coda dark /l/ - Recasens, 2012), English (coda dark /l/ - Sproat & Fujimura 1993), Georgian (clear /l/ in front vowel contexts and coda dark /l/ in back vowel contexts – Robins & Waterson, 1952, Chigogidze, 2011) and Romanian (coda clear /l/ - Recasens, 2012). Unlike dark /l/, clear /l/ lacks an earlier TD retraction (Sproat & Fujimura, 1993) and is therefore not expected to trigger a global organization in coda position. We test our hypothesis on acoustic data. The acoustic effect of global coordination patterns in coda position is a shortening of the vowel in cluster tokens compared to their singleton counterpart.

Methods. A total of 22 native speakers of Russian (5), American English (6), Georgian (5) and Romanian (6) native speakers were recorded producing three repetitions of target singleton-cluster pairs (C)CVL - (C)CVLC with varying front/back vowel contexts embedded in their respective carrier phrase. Formant values (F1, F2, F3) were extracted at the midpoint of the lateral. The darkness degree was determined based on the F2-F1 measure. Two duration measures were considered: (i) vowel + lateral (VL) sequences, and (ii) the interval between the midpoint of the vowel and the midpoint of the lateral (V50-L50) adapting the measure proposed by Durvasula (2023) for onsets. Raw duration measures were normalized dividing the duration by the articulation rate, calculated as the number of phones per second. The duration measure used as a dependent variable in the statistical models was the duration ratio between the cluster and singleton pairs (duration-$VL_{cluster}$ / duration-$VL_{singleton}$). Ratios close to 1 indicate lower degrees of shortening in the cluster token. To compare the degrees of shortening in clusters vs. singletons we compare each language to a hypothetical language (H) which has no shortening. Data for H was generated as a normal distribution of mean=1 and standard deviation equal to the mean standard deviation of the duration ratios found in our data. Linear mixed effects models with *Language* and *Vowel_position* as fixed factors and *Participant* and *Repetition* as random effects with random intercepts were run for each of the two duration measures. An interaction term between *Language* and *Vowel_position* was also included.

Predictions We expect shortening of VL and V50-L50 sequences between clusters and singletons in the case of Russian, English, and back-vowel Georgian codas (dark /l/ in coda). No shortening is expected in Romanian and front-vowel Georgian codas (clear /l/ in coda).

/l/-darkness results show that the degree of lateral darkness is a gradual feature across languages. Russian has the darkest lateral of the four considered languages. Mean values do not significantly differ as a function of vowel position (front: $mean_{F2\text{-}F1}$ = 434, back = 393). English has the second darkest lateral exhibiting a significant difference in F2-F1 values depending on the vowel context: coda /l/ is darker in back vowel contexts ($mean_{F2\text{-}F1}$ = 466) than in front vowel contexts ($mean_{F2\text{-}F1}$ = 585). The third darkest lateral in our data appears in Georgian back vowel contexts ($mean_{F2\text{-}F1}$ = 727). Coda /l/ in Georgian front vowel contexts is much clearer ($mean_{F2\text{-}F1}$ = 1171), confirming the vowel-dependent allophony reported for Georgian. Romanian has clear /l/, independent of the quality of the preceding vowel (front V: $mean_{F2\text{-}F1}$ = 1281; back: $mean_{F2\text{-}F1}$ = 1230)

Duration results only partially confirm our prediction (Fig. 1), and cross-measure differences (VL vs. V50-L50) are found. The VL duration ratio results show that, as expected, English and Russian both show significantly higher degrees of shortening. Romanian shows no differences in shortening compared to the non-shortening hypothetical language in both front and back vowel contexts. Going against our predictions, Georgian exhibits significant shortening in the front vowel context (clear /l/s) and no shortening for back vowel contexts (dark /l/s). The V50-L50 results show the same patterns as well as an additional unpredicted significant shortening for Romanian in front vowel context (i.e., Romanian has significant shortening when compared to the non-shortening hypothetical language only in the context of front vowels). The V50-L50 measure is probably less reliable in our case because of the difficulty of identifying the acoustic boundary between the vowel and the lateral coda, especially in back vowel - dark /l/ and front vowel – clear /l/ sequences.

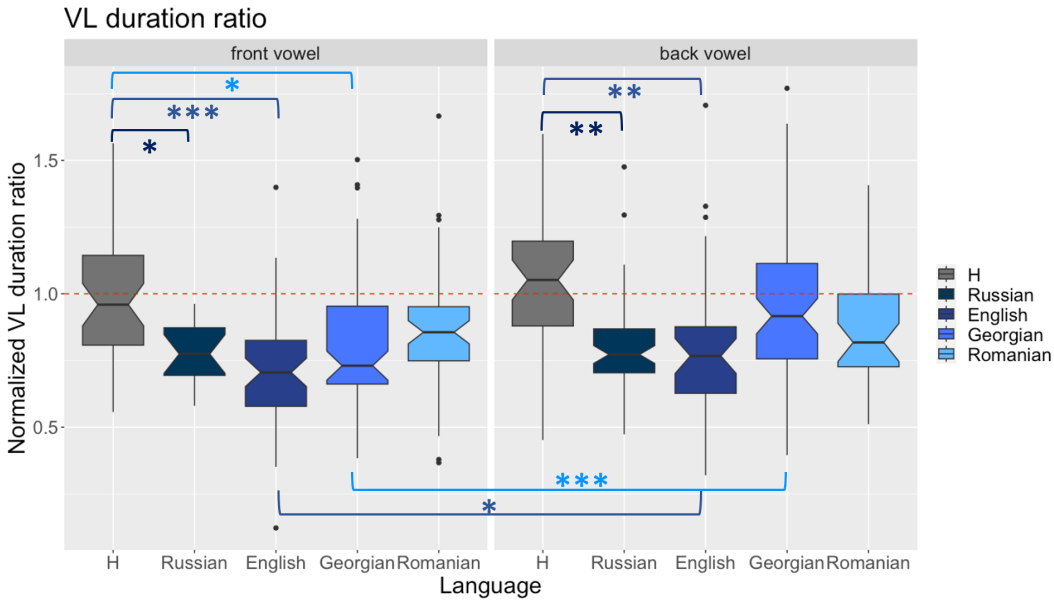


Figure 1: *VL duration as a function of Language and Vowel position,.*
Significance levels indicate differences between the four experimental and the hypothetical non-shortening language H

Discussion. The present paper investigated the hypothesis that global coordination patterns in coda position are triggered by the earlier occurring vocalic gesture present in the production of dark /l/ by comparing four languages that differ in their type of coda lateral. Predictions were confirmed for all languages except Georgian that shows the reverse pattern than the predicted one. One possible explanation for this unexpected pattern is that the degree of darkness could play a role. In our data Georgian dark /l/ is significantly clearer than Russian and English dark /l/. In order to better understand the relationship between /l/ darkness and coda coordination patterns, articulatory data is needed to precisely compare the timing of the articulatory gestures in the lateral coda rime as a function of degree of darkness.

References

Browman CP., & Goldstein L. (1988) Some notes on syllable structure in articulatory phonology. Phonetica; 45(2-4), 140-155

Chigogidze, A. (2011). On the Interaction of Syntax and Phonology in Georgian. MA Thesis.

Durvasula, K. (2023) A simple acoustic measure of onset complexity, ICPhS 2023, 2010-2014.

Marin, S., & Pouplier, M. (2010). Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control, 14*(3), 380–407.

Marin, S. & Pouplier, M. (2014) Articulatory synergies in the temporal organization of liquid clusters in Romanian

Proctor, M., 2009, Gestural characterization of a phonological class: The liquids, Ph.D. Dissertation, Yale University.

Robins, R.H., & Waterson, N. (1952). Notes on the phonetics of the Georgian word. Bulletin of the School of Oriental and African Studies, 141, 55–72.

Recasens, D. (2012) A cross-language acoustic study of initial and final allophones of /l/, Speech Communication 54(3), 368-283.

Sproat R. & Fujimura, O. (1993) Allophonic variation in Enlish /l/ and its implications for phonetic implementation. *Journal of Phonetics. 21(2), 291-311.*

Some Effects of Frame Rate on Gesture Detection in Tongue Ultrasound

Pertti Palo, Steven M. Lulich

Indiana University Bloomington, United States
pertti.palo@taurlin.org, slulich@indiana.edu

Introduction. The analysis of tongue ultrasound data has often been limited to selecting temporal points of interest based on acoustic segmentation and then analysing the corresponding frames by extracting tongue splines. With the advent of reliable automated splining methods (Laporte and Ménard 2018; Wrench and Balch-Tomes 2022) and in the case of holistic image based methods (cf. Palo 2019), we are no longer limited to analyses comparing isolated sample points in time. Instead, we can analyse articulation as a function of time. In time domain analysis the sampling frequency or frame rate of the data becomes a factor which can limit the analysis we are able to perform (Palo and Lulich 2023).
In broad terms, we are interested in understanding how the detectability of speech articulation gestures from tongue ultrasound data is affected by frame rate. According to the Nyquist-Shannon sampling theorem, to detect a signal without aliasing artefacts we need a sampling frequency which is at least double the frequency of the signal. However, to analyze speech timing we also need reasonable time accuracy of the peaks and troughs.
Continuing our recent work (Palo and Lulich 2023), we seek to empirically characterize how the sampling frequency of tongue ultrasound affects automatic detection of articulatory gestures using the Pixel Difference (PD) metric (Palo 2019) and peak detection. PD has so far mainly been defined as the Euclidean distance between consecutive ultrasound frames. Since the Euclidean distance is a special case of a vector lp-norm, we use a selection of additional norms to investigate how frame rate and choice of norm interact in gesture detection.

Materials and Methods. The data is a sample of 174 single-word utterances from a delayed naming experiment. The words were single-syllable lexical English words with a word final plosive ([p, t, k]) and either no onset consonant or an onset of up to three consonants. The data was recorded at about 122 fps in the mid-sagittal plane synchronised with audio (see Experiment 2 in Palo (2019) for details). Data from Speaker P3 are analyzed here.
For this paper, the vector lp-norms are defined as shown in Equation 1. p is the order of the norm and x_i are the first-differences with respect to time of the individual pixel intensities.

$$lp = \begin{cases} \sum_{i=1} \frac{|x_i|}{1+|x_i|}, & p = 0 \\ \sum_{i=1} |x_i|^p, & 0 < p < 1 \end{cases} \quad \text{and} \quad \begin{cases} \sqrt[p]{\sum_{i=1} |x_i|^p}, & 1 \leq p < \infty \\ \max(|x_i|), & p = \infty \end{cases} \tag{1}$$

We chose to use the norms $l0.5$ $l1$, $l2$, and $l5$ to provide a sample around $l1$ and $l2$, which we have used previously, and $l0$ and $l\infty$ because they are the limits of the range of p. Gestures were identified automatically with the function `scipy.signal.find_peaks` from the SciPy software package (Virtanen et al. 2020). We used three parameters – `distance`, `width`, and `prominence` – to tune the peak selection and produce reasonable accuracy in identifying actual gesture peaks. The parameters were selected based on a test set of 10 recordings for norms $l0.5$ $l1$, $l2$, and $l5$.
A conservative lower limit for the gesture interval (parameter `distance`) was estimated from the data of Jacewicz, Fox, and Wei (2010, see Fig. 1). They report a high limit of ~ 6.7 syllables/second for speech rate. Since syllables can be expected to have at least two gestures, we have a lower bound of $t_{lower} = \frac{1}{2*6.7} \approx 0.075$ s for the interval between gestures. This interval length was adapted for downsampling with the formula `distance = t_lower*sampling_frequency` and rounded up. The `width` parameter was chosen as 1 (i.e., a peak with a width of 1 sample or more half way down its prominence value was accepted as valid). Using a higher value would make peak detection deteriorate very fast with downsampling. Finally, the `prominence` value of 0.03 was selected by a step-up process, because using 0.04 excluded peaks in the test set that corresponded to articulatory events.

Results. Our results are illustrated in **Figure 1**. Downsampling causes the number of detected peaks to decline for all norms with $l1$, $l2$, and $l5$ being the most resistant. Similarly, peak position errors increase for all norms while $l1$ and $l2$ behave the best in this respect. All of the error distribution tails cross the $t = 0.075$ s limit already at 41 fps and by 30 fps there more than outliers above the limit for each norm. None of the distribution medians cross the limit before 17 fps.

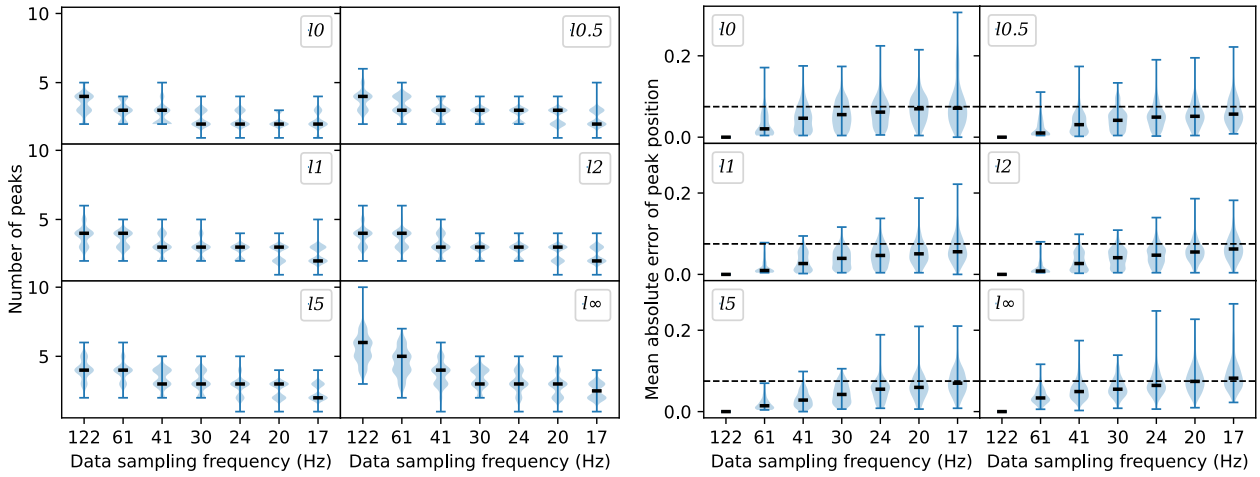


Figure 1: *On the left: ratio of peaks detected. On the right: time accuracy of peak detection compared to the original. Black bars are distribution medians and the dashed line marks the used lower limit between gestures t=0.075 s.*

Discussion. We examined how a particular algorithm for automatic detection of articulation gestures from tongue ultrasound videos is affected by frame rate, and how frame rate interacts with the choice of norm. The ability of $l0$, $l0.5$, and $l\infty$ norms to detect articulatory gestures began to degrade immediately as the frame rate decreased, while the $l1$, $l2$, and $l5$ norms were robust down to 60 fps and degraded more slowly at slower frame rates. The number of detected gestures for $l\infty$ included a high rate of false positives. Finally, the temporal accuracy with which each norm could detect gestures decreased as frame rate decreased, and was subject to especially high error rates for the $l0$, $l0.5$, $l5$, and $l\infty$ norms. Although the present investigation does not identify a specific norm as optimal, both $l1$ or $l2$ are good choices. The $l1$ norm has the advantage of computational efficiency. The $l5$ norm (and higher-order norms) is sensitive to movements limited to within small regions of the ultrasound image where the ultrasound echo is especially strong. This is because the changes in such regions from one frame to the next are large compared with changes in other regions of the image, and these large changes raised to the 5th power dominate the sum in the calculation of the norm. In the same way, the extreme $l\infty$ norm is dominated by the single pixel with the largest change.

References.

Jacewicz, Ewa, Robert Allen Fox, and Lai Wei (2010). "Between-Speaker and within-Speaker Variation in Speech Tempo of American English". In: *The Journal of the Acoustical Society of America* 128.2, pp. 839–850. DOI: 10.1121/1.3459842.

Laporte, Catherine and Lucie Ménard (2018). "Multi-Hypothesis Tracking of the Tongue Surface in Ultrasound Video Recordings of Normal and Impaired Speech". In: *Medical Image Analysis* 44, pp. 98–114. DOI: 10.1016/j.media.2017.12.003.

Palo, P. (2019). "Measuring Pre-Speech Articulation". PhD thesis. Edinburgh: Queen Margaret University.

Palo, P. and S. M. Lulich (Mar. 2023). "Improving Signal-to-Noise Ratio in Ultrasound Video Pixel Difference". In: *The Journal of the Acoustical Society of America* 153.3_supplement, A373. DOI: 10.1121/10.0019222.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Wrench, A. and J. Balch-Tomes (2022). "Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut". In: *Sensors* 22, p. 1133. DOI: doi:10.3390/s22031133.

Laryngeal changes associated with Guttural consonants in Levantine Arabic

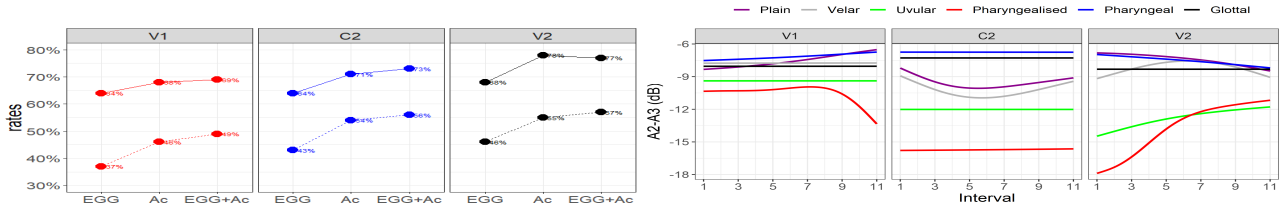
Jalal Al-Tamimi[1]

[1] *Université Paris Cité, CNRS, Laboratoire de linguistique formelle, F-75013 Paris, France*
jalal.al-tamimi@u-paris.fr

Introduction. This study aims at evaluating the role of laryngeal changes associated with guttural consonants in Levantine Arabic. Laryngeal changes are reported in a handful of studies on Arabic guttural consonants; e.g., variable creaky voice using $H1$-$H2$ measure for the variable types of /ʕ/ across Arabic dialects (Heselwood 2007); raised larynx in pharyngealised coronals via videofluoroscopy (F. Al-Tamimi and Heselwood 2011); raised larynx and creaky voice quantified via videofluoroscopy and $H1$-$H2$ in pharyngeals (Heselwood and F. Al-Tamimi 2011); or more recently lower spectral tilt measures of Voice Quality quantifying the degree of tense voice in pharyngealised coronals (J. Al-Tamimi 2017).
Following these findings, we wanted to evaluate laryngeal changes in gutturals using non-invasive techniques. Gutturals are traditionally composed of uvulars /q χ ʁ/ and pharyngeals /ħ ʕ/ (McCarthy 1994) although pharyngealised consonants /tˤ dˤ ðˤ sˤ/ and glottals /ʔ h/ are also included due to the constriction location being somewhere in the lower vocal tract. Recently, J. Al-Tamimi and Palo (2023) provided an empirical evidence that gutturals share articulatory similarities in how they impact on the tongue contours quantified via Ultrasound Tongue Imaging (UTI). The results also suggested that gutturals (uvular, pharyngealised and pharyngeal) show a potential for a raised larynx posture due to a variable degree of tongue root retraction. However, and due to the difficulty in quantifying laryngeal changes from UTI alone, this study uses a combined acoustic and Electroglottography (EGG) measures to quantify the (dis-)similarities between gutturals with respect to systematic laryngeal changes. We ask the question whether gutturals show a systematic larynx raising, when compared to plain coronals (or velars and glottals) and whether there is a systematic spectral tilt lowering, which could be indicative of a more creaky or tense voice quality. Following the predictions of the Laryngeal Articulator Model; LAM (Esling et al. 2019) and the findings from J. Al-Tamimi and Palo (2023), we expect gutturals to show systematic differences to non-gutturals but to show a more gradient larynx height and constriction (quantified via Closed Quotient; CQ and Peak Increase in Contact; PIC). We also expect them to show a gradient spectral tilt decrease as quantified via the various acoustic measures of voice quality.

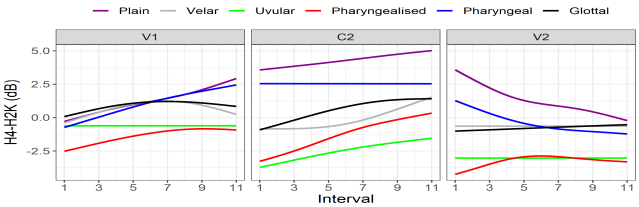
Methods. Ten Levantine Arabic Urban speakers (5 males, 5 females), aged 25-45 were recorded using synchronised UTI, 2-channel EGG system, and audio recordings through a multichannel breakout. They were instructed to produce a list of items in a /ʔV:CV:/ frame, with all possible consonants in Levantine Arabic and the three symmetric long vowels /i: a: u:/; see more details of corpus and data collection in J. Al-Tamimi and Palo (2023). After transcribing and force-aligning the data using Arabic WebMUS (J. Al-Tamimi, Schiel, et al. 2022), the data were processed. Larynx Height (LH) was extracted from the 2nd channel and contact measures were automatically quantified using EGGWorks. EGG and acoustic measures of voice quality were obtained at 11-intervals from VoiceSauce, with $f0$ and formant measures being speaker-adapted using Praat settings. For the EGG measures, we obtained LH, CQ (using the Hybrid method); the PIC and PDC (Peak Decrease in Contact); the Closing and opening slopes and ratio (Kuang and Keating 2014). For acoustic measures, we extracted all measures described in the psychoacoustic model of voice quality (Garellek et al. 2016) in addition to the three high-frequency amplitude measures used in J. Al-Tamimi (2017) to quantify the impact of an epilaryngeal constriction on the voice source. We used Random Forest (RF), a Machine Learning classifier, and dynamic information to evaluate the impact of gutturals on the VCV sequence. We look next at the classification results and the patterns observed for the top three measures identified by the various RFs we ran.

Results. Figure 1a shows classification results. When combining EGG and acoustic measures (EGG+Ac), an overall improvement in classification accuracy is observed especially within V2. This is indicative of a progressive coarticulatory impact of voice quality. When comparing gutturals to non-gutturals (2 classes; solid lines), the rates are around 78%; when looking at the six classes (dashed lines), the rates are much lower and are close to the 60%. The relatively low rates

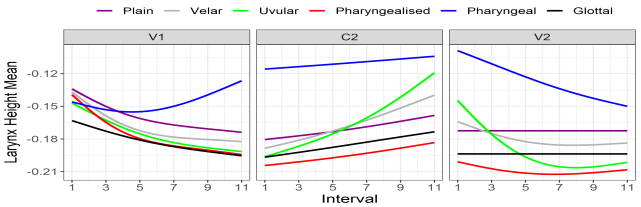


(a) Classification results; solid = 2 classes

(b) *A*2*-*A*3*



(c) *H*4*-*H*2k*



(d) Larynx Height

are indicative of a secondary role for voice quality. Figures 1b, 1c and 1d show the top three predictors used by most of the RF. Dynamic changes throughout the VCV sequence show values to have an overall ⇓ in V1, ⇑/⇓ in C2 and ⇑ in V2. Interestingly, uvular and pharyngealised consonants show ⇓*A*2*-*A*3* and ⇓*H*4*-*H*2k* indicative of a steeper and an abrupt change in the Slope around 2kHz and 3kHz, which can be correlated with a more constricted glottis; pharyngeal consonants show the highest values indicating a potential for a more constricted epilarynx. Pharyngeal consonants have a clear ⇑LH, throughout the VCV sequence, which is more marked at the C2; pharyngealised consonants show a pattern of ⇓ in V1, ⇑ in C2 and ⇓ in V2, while uvular consonant follows a similar pattern, but shows a ⇑LH within the release.

Discussion. These results confirm that gutturals share specific common laryngeal components. Pharyngeal consonants are inherently produced with a ⇑Larynx, which yields a constricted epilarynx, following LAM. Pharyngealised consonants on the other hand show ⇓Larynx, which increases towards the release, but with a more abrupt and constricted glottis. Uvular consonant seems to share a ⇑Larynx *and* abrupt and constricted glottis similar to pharyngealised. The results of this study provide an empirical evidence that gutturals show systematic laryngeal changes not often quantified in the literature.

References.

Al-Tamimi, Feda and Barry Heselwood (2011). "Nasoendoscopic, Videofluoroscopic and Acoustic Study of Plain and Emphatic Coronals in Jordanian Arabic". In: *Instrumental Studies in Arabic Phonetics*. Ed. by B. Heselwood and Z. Hassan. John Benjamins, pp. 165–191. DOI: 10.1075/cilt.319.

Al-Tamimi, Jalal (2017). "Revisiting Acoustic Correlates of Pharyngealization in Jordanian and Moroccan Arabic: Implications for Formal Representations". In: *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8.1, pp. 1–40. DOI: 10.5334/labphon.19.

Al-Tamimi, Jalal and Pertti Palo (2023). "Dynamics of the Tongue Contour in the Production of Guttural Consonants in Levantine Arabic". In: *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*. Ed. by Radek Skarnitzl and Jan Volín. Prague, Czech Republic (7-11 August 2023): Guarant International, pp. 2095–2099.

Al-Tamimi, Jalal, Florian Schiel, Ghada Khattab, Navdeep Sokhey, Djegdjiga Amazouz, Abdulrahman Dallak, and Hajar Moussa (2022). "A Romanization System and WebMAUS Aligner for Arabic Varieties". In: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), l' European Language Resources Association (ELRA), Licensed under CC-BY-NC-4.0*. Marseille, 20-25 June 2022, pp. 7269–7276.

Esling, John, Scott Moisik, Allison Benner, and Lise Crevier-Buchman (2019). *Voice Quality: The Laryngeal Articulator Model*. 1st ed. Cambridge University Press. DOI: 10.1017/9781108696555.

Garellek, Marc, Robin Samlan, Bruce R. Gerratt, and Jody Kreiman (2016). "Modeling the Voice Source in Terms of Spectral Slopes". In: *The Journal of the Acoustical Society of America* 139.3, pp. 1404–1410. DOI: 10.1121/1.4944474.

Heselwood, Barry (2007). "The 'Tight Approximant' Variant of the Arabic 'ayn". In: *Journal of the International Phonetic Association* 37.01, p. 1. DOI: 10.1017/S0025100306002787.

Heselwood, Barry and Feda Al-Tamimi (2011). "A Study of the Laryngeal and Pharyngeal Consonants in Jordanian Arabic Using Nasoendoscopy, Videofluoroscopy and Spectrography". In: *Instrumental Studies in Arabic Phonetics*. Ed. by B. Heselwood and Z. Hassan. John Benjamins, pp. 101–128. DOI: 10.1075/cilt.319.

Kuang, Jianjing and Patricia Keating (2014). "Vocal Fold Vibratory Patterns in Tense versus Lax Phonation Contrasts". In: *Journal of the Acoustical Society of America* 136.5, pp. 2784–2797. DOI: 10.1121/1.4896462.

McCarthy, John (1994). "The Phonetics and Phonology of Semitic Pharyngeals". In: *Phonological Structure and Phonetic Form*. Ed. by Patricia Keating. Cambridge: Cambridge University Press, pp. 191–233. DOI: 10.1017/CBO9780511659461.012.

Vocal Motor Control During Exposure to Oscillating Pitch Changes

Rita Bishai, Jeffery A. Jones, Nichole E. Scheerer

Psychology Department, Wilfrid Laurier University, Waterloo, Ontario N2L 3C5, Canada

bish9250@mylaurier.ca, jjones@wlu.ca, nscheerer@wlu.ca

Introduction. Feedback and feedforward motor control systems are used to produce and regulate the controlled movements associated with speech (Guenther 1995; Guenther et al. 2006). The feedforward system relies on stored representations that map the relationship between articulator movements and their associated speech sounds (Guenther & Vladusich 2012). Activation of these representations plays an important role in initiating the production of specific speech targets. Conversely, the feedback system continually compares the auditory and somatosensory feedback produced during speech production with target representations from the feedforward system (Guenther 1995; Guenther et al. 2006). This comparison allows the feedback system to identify and initiate corrections for speech errors. It is widely accepted that these two systems work in combination with each other to produce fluent speech, however, the relative weighting of these systems is thought to differ across speakers and contexts.

To study the use of these systems, frequency altered feedback (FAF) paradigms are used to alter the fundamental frequency (F0) of a speaker's voice to simulate a speech error. Changes in F0 production in response to these manipulations can then be measured (Elman 1981; Jones & Keough 2008; Larson et al. 1995; Scheerer and Jones 2012). Typically, speakers will compensate for changes in their auditory feedback, with the magnitude of this response thought to be indicative of the extent to which a speaker is using their feedback system (Scheerer and Jones, 2012). Interestingly, Scheerer and Jones (2012) found that participants with more variable voices produced larger compensatory responses to FAF, suggesting that they rely more heavily on their feedback system. As such, larger compensatory responses, and increased weighting of feedback control, have been linked to vocal variability (Parrell & Houde 2019).

FAF can also be used to examine sensorimotor learning by shifting a speaker's F0 in a predictable manner (Jones and Keough 2008; Keough et al. 2013). Under these circumstances, speaker's not only compensate for the change in F0, but they also continue to compensate when the alteration is removed. The persistence of the compensatory response suggests that the deviant feedback was used to update the feedforward representation based on the new relationship between articulator movements and their associated speech sounds. Notably, vocal variability has also been linked to the amount of sensorimotor learning that takes place. Jones and Keough (2008) found that singers, who were less variable, showed greater sensorimotor learning compared to non-singers.

While previous studies have established a link between vocal variability, compensatory responses, and sensorimotor learning, these studies are all correlational in nature. The current study aims to further explore these relationships by experimentally inducing variability in speakers' voices. Given previous findings, we expect that by introducing variability into a participant's voice it will promote increased reliance on feedback control, resulting in larger compensatory responses and reduced sensorimotor learning.

Methods. Fifty-one Canadian-English speaking adults between the ages of 17 and 56 years old (12 male, 39 female) participated in this study. Participants were asked to say the vowel sound /a/ while their voice was fed back to them in real time. Participants produced 180 vocalizations across 2 blocks. During the "control" block, participants' F0 was unaltered for the first 25 trials, shifted ±100 cents for 50 trials, then unaltered for the remaining 15 trials. For the "oscillation" block, participants' auditory feedback was oscillated in the form of a 25-cent triangle wave at a rate of 4 Hz using 1 cent shift steps. In this block, the first 5 trials were unaltered, then 20 trials were oscillated, then while still oscillating 25 cents, their auditory feedback was also shifted ±100 cents for 50 trials, finally, their auditory feedback was unaltered for the final 15 trials. The shifted trials in the control and oscillation blocks were always shifted in opposite directions and the block order was randomized. Thus each participant was exposed to one control and one oscillation block, with one shifted in the positive and the other in the negative direction.

For each vocalization, the median F0 of the first 80 and 1500 ms of the vocalization was calculated, median 80 and median 1500, respectively. These values were calculated in cents by normalizing the F0 of the vocalization in Hz to the averaged median 1500 of the five baseline trials at the start of each block. Since neural processing delays prevent the auditory feedback resulting from a vocal motor command from being processed for at least 100 ms (Burnett et al. 1997), the median 80 value provides an index of the F0 at which the vocalization was initiated, before

auditory feedback was available. On the other hand, median 1500 values provide an index of the F0 of the vocalization once auditory feedback becomes available. Median 80 and 1500 values were then averaged into trial groups, each group containing 5 trials. Trial groups were organized into 4 categories, baseline (1), pre-shift (2-5), shift (6-15), and test (16-18). Prior to analysis, median 80 and 1500 values were collapsed across shift direction (up/down) by multiplying median values for the up conditions by -1. This resulted in two final conditions: control and oscillation.

Results. *Median 80*. A repeated measures analysis of variance (RM-ANOVA) revealed a significant main effect of trial group, $F(17,850) = 9.13$, $p < .001$, $\eta^2p = 0.154$, a marginal main effect of block, $F(1,50) = 3.67$, $p = .061$, $\eta^2p = 0.068$, and a significant interaction between trial group and block, $F(17, 850) = 2.40$, $p = .001$, $\eta^2p = 0.046$ (see Figure 1). *Median 1500.* A RM-ANOVA revealed a significant main effect of trial group, $F(17,850) = 11.62$, $p < .001$, $\eta^2p = 0.226$, and a significant interaction between trial group and block, $F(17, 850) = 6.64$, $p < .001$, $\eta^2p = 0.068$ (see Figure 1). The main effect of block was not significant.

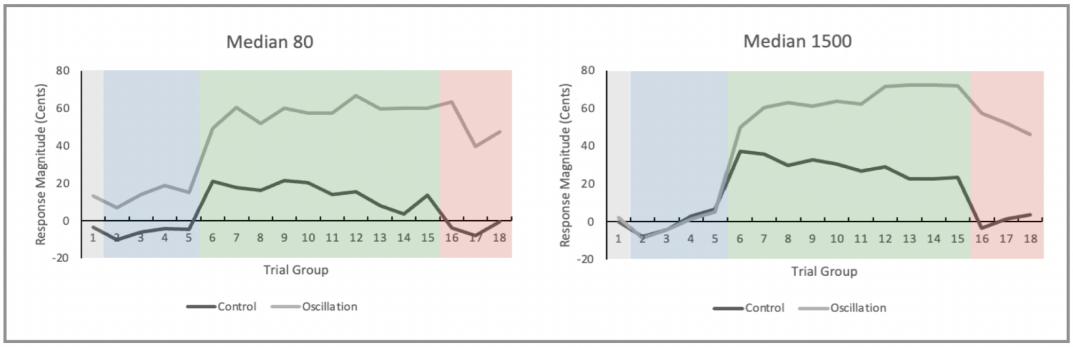


Figure 1: Control (black line) and oscillation (grey line) F0 values plotted for median 80 and median 1500. Trial groups are highlighted as follows: baseline (grey), pre-shift (blue), shift (green), test (red).

Discussion. This study examined the effects of induced vocal variability on sensorimotor learning and compensatory responses when speakers were exposed to FAF. The results indicated that when speakers' auditory feedback was oscillated, introducing variability to their speech, participants produced larger compensatory responses to predictable manipulations to their auditory feedback. Further, in conditions of increased variability, when speakers' auditory feedback was returned to its unaltered state, speaker's showed evidence of greater sensorimotor learning.

While past research has shown that individuals who are more variable in their speech production produce larger compensatory responses to FAF (Scheerer and Jones 2012), the current study extends these findings by demonstrating that inducing variability in a speaker's auditory feedback causes them to produce larger compensatory responses to FAF, relative to when they are exposed to FAF while their auditory feedback is otherwise unaltered. This suggests that under conditions of increased variability, speakers rely more on their auditory feedback to monitor and correct for errors in their speech. At the same time, this increased reliance on feedback control makes them more susceptible to FAF manipulations, and thus larger compensatory responses are observed. Importantly, these findings suggest that the weighting of feedback and feedforward control can be dynamically altered to accommodate temporary changes in the reliability of the feedforward system.

The results of this study also suggest that greater sensorimotor learning occurs under conditions of increased vocal variability. This finding was unexpected as past research has suggested singers, who are more stable, show more sensorimotor learning when exposed to FAF, relative to more variable non-singers. Future research will continue to explore the relationship between vocal variability and sensorimotor learning.

References.

Burnett, T. A., Senner, J. E., & Larson, C. R. (1997). *Journal of Voice*, *11*(2), 202–211.
Elman, J. L. (1981). *The Journal of the Acoustical Society of America*,70(1), 45–50.
Guenther, F. H. (1995). *Psychological Review*, *102*(3), 594–621.
Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). *Brain and Language*, *96*(3), 280–301.
Guenther, F. H., & Vladusich, T. (2012). *Journal of Neurolinguistics*, *25*(5), 408–422.
Jones, J. A., & Keough, D. (2008). *Experimental Brain Research*, *190*(3), 279–287.
Keough, D., Hawco, C., & Jones, J. A. (2013). *BMC Neuroscience*, *14*(1), 25.
Larson, C. R., Carrell, T. D., Senner, J. E., Burnett, T. A., & Nichols, L. L. (1995). *Vocal fold physiology: Voice quality control*, 321-31.
Parrell, B., & Houde, J. (2019). *Journal of Speech, Language, and Hearing Research*, *62*(8S), 2963–2985.
Scheerer, N. E., & Jones, J. A. (2012). *Neuroscience Letters*, *529*(2), 128–132.

Past sensory prediction error and motor compensation asymmetrically mediate speech motor control

Yuhan Lu[1], Tingting Wang[2], Xing Tian[1]

[1]New York University Shanghai, [2]Minerva University

yuhan.lu@nyu.edu, tingting.wang@uni.minerva.edu, xing.tian@nyu.edu

Introduction. How the brain learns and controls vocal production is an important question to neuroscientist and linguistics. It has been proposed that infants' babbling and adults adopting new dialects hinge on the integration between motor efference copy and its auditory and somatosensory feedback. When discrepancies arise between these signals, the brain adjusts subsequent vocal production. However, it is ongoingly debated what prior experiences are retained to shape future production. One hypothesis proposes learning from past sensory prediction errors derived from differences between prior motor efference copy and sensory feedback to update the internal model (Houde and Nagarajan, 2011). Another hypothesis suggests that rather than the sensory prediction error itself, its subsequent corrective movements are learned and integrated into the feedback motor program for future production (Tourville and Guenther, 2011). Previous study on auditory perturbation has found that the previous motor compensatory response to perturbation can adapt to the following vocal production without perturbation (Hantzsch et al., 2022); However, the two stages of speech motor control, i.e., sensory prediction error and motor compensation, are highly coupled, it is difficult to delineate the influence of these two stages on subsequent vocal production. In our current study, we employed a lengthy sequence to quantify randomly-applied perturbations and their compensations, and thus to measure the serial bias of previous sensory prediction errors and motor compensations on present vocal adjustments.

Methods. Fifteen female participants with self-reported normal hearing and speech were recruited in the study. The experiments had 960 trials in three days, with 320 trials each. In each trial, subjects were asked to produce /a/ for 2.5 seconds when a green cross was presented on the screen. After vocalization onset, a pitch perturbation was introduced randomly between 0.8-1.2 s and lasted 500 ms (**Fig. 1A**). The pitch perturbation direction (i.e., upward and downward) and amount (i.e., 60, 180, 300 cents) were randomly assigned to 960 trials for each subject. Next trial was presented after a silent interval randomized between 1-2 s (**Fig. 1B**).

Subjects' pitch traces in each trial (at 100-Hz sampling rate) were measured using automatic Praat-based script ProsodyPro (Xu, 2013). The baseline pitch was calculated by taking the mean of pitch trace in the -400 to -200 ms before perturbation onset. Each pitch trace was then normalized and converted from Hz to cent using its average baseline pitch. The perturbation and compensation amounts were measured by taking the averaged mean in the 50-80 ms and in the 150-250 ms time window, respectively, minus mean baseline pitch. Compensation amounts for all trials and all subjects were corrected by subtracting the grand average. Next, we separately quantified the systematic biases of current compensation amount (C_t) by the perturbation (P_{t-1}) and compensation amounts (C_{t-1}) of vocal production presented on the 31 preceding trials (**Fig. 1C**). For the immediately preceding (1-back) trial, current compensatory response was expressed as a function of the difference between previous (P_{t-1}) and current perturbation amounts. For positive values of this difference, the previous perturbated pitch was higher than the current perturbated pitch. Consequently, data points in the scatter plot that had x- and y- value in the same sign indicated that the current compensation was in the direction of the previous perturbation. The same principle was also applied for the 1-back motor-compensation bias.

We built four generalized linear mixed-effect models to quantitatively test the 1-back serial bias on current compensation. Model 1 assumed that the current compensation had no history influence but was only responded to the perturbation (P_t), models 2 and 3 assumed, respectively, on the basis of current perturbation, previous perturbation and previous compensation influences current compensation, and Model 4 assumed both previous perturbation and compensation taken together influence the current compensation. Model performance was evaluated using the ΔAIC, which assessed the goodness of fit of the model with a smaller value indicating better performance. Regression coefficient of each predictor in the full model was accessed by the t-statistics against the null hypothesis test that the coefficient was equal to 0. For multiple comparison, a false discovery rate correction was applied.

Results. We analyzed the dependence of current compensatory response, separately, on difference of current and perturbation in the previous one trial (1-back; P_{t-1} - P_t) and on 1-back compensation (C_{t-1}) (**Fig. 2C**). Therefore, we had current compensation as the function of 1-back perturbation and the function of 1-back compensation for each subject (**Fig. 2D**). We found that current compensation was systematically attracted towards both 1-back perturbation and 1-back compensation after averaging across subjects. These attractions yielded derivative-of-Gaussian-shaped curve ($p = 0.0001$ for both, permutation test), similar to previous reports of visual serial dependency effect (Fischer and Whitney, 2014;

Liberman et al., 2014). We build four models to assess the contribution of 1-back perturbation and 1-back compensation to the current compensation. Model 1 assumed that current compensation only depended on the current perturbation, which was a base model without considering serial bias. Model 2 and 3 additionally considered contribution of 1-back perturbation and 1-back compensation, respectively. Model 4 took account of contributions from all, i.e., current perturbation, 1-back perturbation, and 1-back compensation. Models 2-4 were evaluated against the base Model 1 by comparing the Akaike information criterion (AIC). It showed that Models 2-4 considering 1-back history outperformed the base Model 1 (Model 1: ΔAIC = 36,625.33; Model 2: ΔAIC = 34,737.13; Model 3: ΔAIC = 34,692.02; Model 4: ΔAIC = 34,686.94; **Fig. 2E**, upper panel), suggesting influence of 1-back perturbation and compensation on current compensation. Critically, inset figure further showed that Model 3 was better than Model 2, suggesting that 1-back compensation contributed more than 1-back perturbation. Moreover, we compared the regression coefficient of these three predictors in the full Model 4 (**Fig. 2E**, lower panel), to verify the contribution weights of 1-back factors. It showed positive coefficients of 1-back perturbation (0.02 ± 0.009, $t(12,532) = 2.66$, p = 0.008, one-sample t-test) and 1-back compensation (0.06 ± 0.009, $t(12,532) = 7.23$, $p = 5 \times 10^{-13}$, one-sample t-test), which confirmed attractive serial bias, with stronger in 1-back compensation than 1-back perturbation. These results suggested that compensatory response attracted towards both short-term sensory-prediction-error history and motor-compensation history, but the motor-compensation serial bias was stronger.

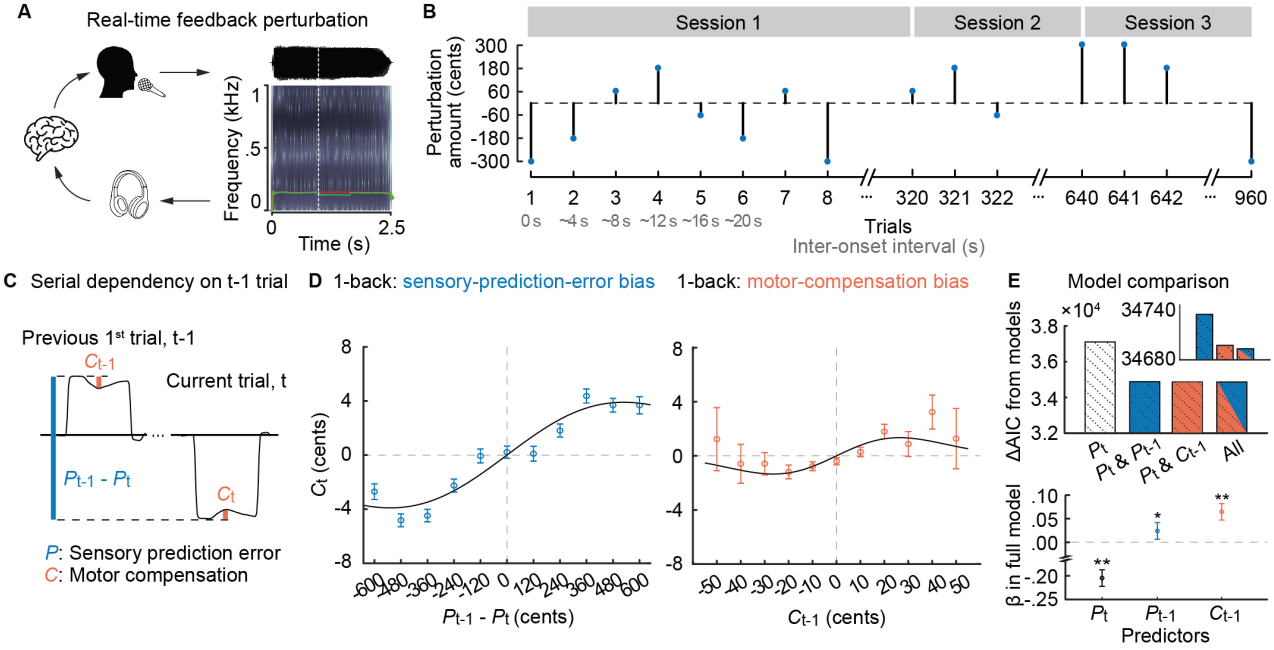


Figure 1: *Study paradigm (A-B) and serial dependency of past sensory-prediction-error and motor-compensation histories on current compensation (C-E).*

Discussion. We investigated the relationship between current compensatory responses and their dependence on preceding perturbation and compensation in a 1-back context. we observed that current compensation exhibited a notable attraction towards both the 1-back perturbation and 1-back compensation, and emphasized a stronger attraction serial bias in motor-compensation history than sensory-prediction-error history. The novelty of our study lies in introducing a fresh paradigm to investigate speech motor control, particularly in characterizing the impact of past production history. It has long-been focused on understanding why compensatory responses to auditory perturbations can vary between following or opposing patterns. Past studies primarily yielded qualitative insights into factors such as confidence levels in prior auditory feedback and cortical sensitivity. However, our current research offers a new explanation: the nature of subsequent compensatory responses is intricately tied to previous auditory-motor integration and motor compensation.

References

Fischer, J., Whitney, D., 2014. Serial dependence in visual perception. Nat Neurosci 17, 738-743. http://dx.doi.org/10.1038/nn.3689

Hantzsch, L., Parrell, B., Niziolek, C.A., 2022. A single exposure to altered auditory feedback causes observable sensorimotor adaptation in speech. Elife 11. http://dx.doi.org/10.7554/eLife.73694

Houde, J.F., Nagarajan, S.S., 2011. Speech production as state feedback control. Front Hum Neurosci 5, 82. http://dx.doi.org/10.3389/fnhum.2011.00082

Liberman, A., Fischer, J., Whitney, D., 2014. Serial dependence in the perception of faces. Curr Biol 24, 2569-2574. http://dx.doi.org/10.1016/j.cub.2014.09.025

Tourville, J.A., Guenther, F.H., 2011. The DIVA model: A neural theory of speech acquisition and production. Lang Cogn Process 26, 952-981. http://dx.doi.org/10.1080/01690960903498424

Xu, Y., 2013. ProsodyPro—A tool for large-scale systematic prosody analysis. Laboratoire Parole et Langage, France.

The Acoustics of Vowel Sequences in Five Romance Languages

Johanna Cronenberg[1], Lori Lamel[2], Ioana Chitoran[1]

[1]*Université Paris Cité,* [2]*Laboratoire Interdisciplinaire des Sciences du Numérique*

johanna.cronenberg@u-paris.fr, lamel@lisn.fr, ioana.chitoran@u-paris.fr

Introduction. This study is concerned with the acoustics of the vowel sequences /ia/ and /io/ in five Romance languages. In French and Italian, these sequences are typically produced as diphthongs while hiatus is preferred in Portuguese (Chitoran and Hualde 2007). In Spanish (Herrero de Haro and Alcoholado Feltstrom 2023; Hualde and Prieto 2002) and Romanian (Chitoran 2002; Marin 2004), both diphthongs and hiatuses can occur. The classification of /ia/ and /io/ as either diphthongs or hiatuses in these languages is usually based on phonological criteria or on small samples of carefully read words. The aim of this study is therefore to provide details about the acoustic configuration of the vowel sequences via an analysis of their realisation in large corpora of fluent speech. In this preliminary investigation we focus on formant slope and curvature as an indication of the distinction between diphthongs and hiatuses. Hiatuses should be characterised by a steeper formant slope and greater curvature compared to diphthongs, given that the transition between the two vocalic elements is sharper in hiatuses than in diphthongs (Aguilar 1999).

Methods. The data used for this study came from corpora of European radio shows in French, Italian, Portuguese, Romanian, and Spanish, which aired between 1992 and 2012 (see Vasilescu et al. 2020, for details on the size of the corpora). These recordings were processed through an ASR algorithm, i.e. they were orthographically transcribed and phonemically segmented via forced alignment with word-context independent phone models (Lamel and Gauvain 1992; Adda-Decker and Lamel 1999). All sequences of /i/ followed by /a/ or /o/ in non-word-final and non-word-initial position were extracted from the audio and segmentation files. Table 1 shows the number of available vowel sequences per language. We measured the first two formants using the *forest* algorithm from the R package `wrassp`, and Lobanov-normalized them. For every F1 and F2 trajectory, a discrete cosine transform was calculated and the coefficients DCT-1 (slope) and DCT-2 (curvature) were extracted using the R package `dtt`. These coefficients were then used as the dependent variables in four mixed models, one for each formant $\times$ DCT coefficient combination, with `lmerTest`. The models further included language (five levels), vowel sequence (two levels: /ia/, /io/), and their interaction as fixed factors, as well as random intercepts for word (13,259 levels) and audio file (234,066 levels). Finally, post-hoc comparisons were estimated with `emmeans`.

Results. Almost all post-hoc comparisons between the languages are highly significant. Figure 1 shows F1 and F2, reconstructed using the estimated marginal means for DCT-1 and DCT-2 in an inverse discrete cosine transform, i.e. these can be viewed as formants that consist only of the typical slope and curvature for each language and vowel sequence. Portuguese shows the steepest F2 slope for both /ia/ and /io/ as well as steep F1 slopes, which is consistent with the rapid transition between the two elements of a hiatus. French, on the other hand, shows shallow F2 trajectories for both vowel sequences, as well as a shallow F1 trajectory for /ia/, which aligns with the expectation that French speakers produce these sequences as diphthongs. The F2 trajectories of Spanish are very shallow, while the F1 trajectories are rather steep. The opposite is true of Romanian (shallow F1, steep F2). The Italian formants are particularly curvy compared to those of the other four languages, while their steepness is similar to that of the Romanian formants.

Discussion. This study is the first to our knowledge that provides a cross-linguistic comparison of the formant configuration of /ia/ and /io/ on the basis of a large amount of data extracted from fluent speech. Our preliminary analysis has shown that the spectrum between diphthongs and hiatuses, which has been claimed to exist based on a phonological-historical point of view and duration measurements (Chitoran and Hualde 2007), is not quite as clear in acoustic formant data. While Portuguese and French show the expected patterns, we did not identify variation in Spanish and Romanian that would have indicated that both diphthongs and hiatuses are produced. Spanish, however, has recently been shown to

favour diphthongs (Herrero de Haro and Alcoholado Feltstrom 2023), and the partially monophthongal formants in our Romanian data might result from the large proportion of vowel sequences preceded by /ts/ or /tS/ (53.1% for /io/, 38.4% for /ia/) which can lead to monophthongisation (Chitoran 2002). In future investigations, we plan to include factors which are likely to influence the production of these vowel sequences, such as lexical stress, position within the word, and phonetic context. We further aim to broaden our analyses to include more measures, in particular duration which has been shown to be highly relevant in the acoustic and perceptual distinction between diphthongs and hiatuses.

Table 1: *Count of /ia/ and /io/ sequences analysed in the five languages.*

Language	**/ia/**	**/io/**	
French	6,400	7,418	**13,818**
Italian	24,880	30,906	**55,786**
Portuguese	10,008	429	**10,437**
Romanian	29,334	13,540	**42,874**
Spanish	34,045	82,889	**116,934**
	104,667	**135,182**	

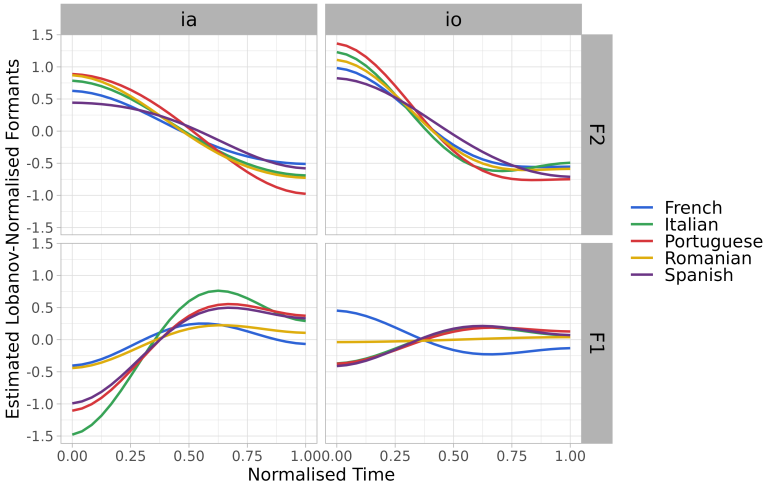


Figure 1: *Lobanov-normalised F1 (bottom) and F2 (top) over normalised time, aggregated by vowel sequence (/ia/ left, /io/ right) and language (colour-coded), reconstructed using the estimated marginal means of DCT-1 and DCT-2.*

References.

Adda-Decker, Martine and Lori Lamel (1999). "Pronunciation Variants across System Configuration, Language and Speaking Style". In: *Speech Communication* 29, pp. 83–98.

Aguilar, Lourdes (1999). "Hiatus and Diphthong: Acoustic Cues and Speech Situation Differences". In: *Speech Communication* 28, pp. 57–74.

Chitoran, Ioana (2002). "A Perception-Production Study of Romanian Diphthongs and Glide-Vowel Sequences". In: *Journal of the International Phonetic Association* 32.2, pp. 203–222. DOI: 10.1017/S0025100302001044. (Visited on 04/25/2023).

Chitoran, Ioana and José Ignacio Hualde (2007). "From Hiatus to Diphthong: The Evolution of Vowel Sequences in Romance". In: *Phonology* 24.1, pp. 37–75. DOI: 10.1017/S095267570700111X. (Visited on 04/25/2023).

Herrero de Haro, Alfredo and Antonio Alcoholado Feltstrom (2023). "Anti-Hiatus Tendencies in Spanish: Rate of Occurrence and Phonetic Identification". In: *Linguistics*, pp. 1–26. DOI: 10.1515/ling-2021-0228. (Visited on 12/12/2023).

Hualde, José Ignacio and Mónica Prieto (2002). "On the Diphthong/Hiatus Contrast in Spanish: Some Experimental Results". In: *Linguistics* 40.2, pp. 217–234. DOI: 10.1515/ling.2002.010. (Visited on 04/25/2023).

Lamel, Lori and Jean-Luc Gauvain (1992). "Continuous Speech Recognition at LIMSI". In: *Proceedings of the Final Review of the DARPA ANNT Speech Program*. Stanford, pp. 1–7.

Marin, Stefania (2004). "Complex Nuclei in Articulatory Phonology: The Case of Romanian Diphthongs". In: *Proceedings of the 34th Linguistic Symposium on Romance Languages*. Salt Lake City, pp. 161–177.

Vasilescu, Ioana, Yaru Wu, Adèle Jatteau, Martine Adda-Decker, and Lori Lamel (2020). "Alternances de Voisement et Processus de Lénition et de Fortition: Une Étude Automatisée de Grands Corpus En Cinq Langues Romanes". In: *Traitement Automatique des Langues* 61.1, pp. 11–36.

Does Cross-Word Resyllabification Modify Word Cohesion in French?

Alice Yildiz, Anne Hermes, Cécile Fougeron

Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), Paris, France
alice.yildiz@sorbonne-nouvelle.fr, anne.hermes@sorbonne-nouvelle.fr,
cecile.fougeron@sorbonne-nouvelle.fr

Introduction. Defining the phonetic reality of linguistic units such as a phoneme, a syllable or an intonational phrase has been a longstanding endeavor in phonetics. Here we focus on a unit often left aside: the word. If the word has a phonetic reality, then its internal component should show a specific cohesion, and/or its boundaries should be marked somehow in the speech signal. In the present study, we tackle this question by looking at *enchaînement*, a process in which word boundaries are said to be blurred by a resyllabification process of the coda consonant of the word 1 (W1) with the vowel starting the following word 2 (W2): e.g., W1 <petite> + W2 <ami> /pə.tit#a.mi/ > [pə.ti.ta.mi] (with "." and "#" for syllable and word boundaries, respectively). This sandhi process thus creates a misalignment between syllable and word boundaries with the underlying final coda of W1 being resyllabified as the onset of W2. However, several studies have shown that the *enchaînement* consonant is not like a true onset) and that some of the word-specific characteristics of W1 are maintained even if there is resyllabification, so that resyllabified /V_1.C#V_2/ sequences ([ita] in <petite> + <amie>) are distinct from sequences with a true word-initial onset /V_1.#CV_2/ ([ita] in <petit> + <tamis>) (e.g., Rousselot 1901; Yersin-Besson & Grosjean 1996; Fougeron 2007; Bagou et al. 2009). Indeed, the consonant, which is underlyingly a coda of W1, is found to be generally shorter than in a true word-initial onset position (e.g., Fougeron 2007; Bagou et al. 2009). Furthermore, V_1 which is underlyingly in the closed word-final syllable of W1 (V_1C#) is found to have a higher F1 than in open word-final syllables (V_1#C) (Fougeron 2007; L'Esperance 2015). We aim to investigate further whether the *enchaînement* process in French modifies the word cohesion. This study follows on this question by (i) confirming that durational and spectral cues on C and preceding vowel indeed distinguish the underlying word final VC# in enchainment context from a V_1#C sequence; and (ii) by investigating what could be another aspect of the internal cohesion of the underlying word final V_1C#: the temporal stability of the acoustic lag between V and C. If we assume that gestural coherence is stronger within word and syllable units (Byrd 1996; Cho 2001; Yoon et al. 2011; cf. Nam & Saltzman 2003; Mücke 2018) then we expect less stability in the acoustic lag between V_1 and C lose their syllabic cohesion with resyllabification.

Method. To test this, we use data extracted from two French data sets containing the same V_1CV_2 sequences in both enchaînement (V_1.C#V_2) and word-initial conditions (V_1#.CV_2) (e.g., "petit tamis" /pə.ti#ta.mi/ vs. "petite amie" /pə.tit#a.mi/). In the first corpus, the sequences include: /ap#a/ vs. /a#pa/ and /aʁ#a/ vs. /a#ʁa/ sequences, as in e.g., "Le pape a donné un discours à Rome." – '*The Pope gave a speech in Rome*.' These sentences were recorded by five French female speakers (22 to 28 years), and repeated 45 times (for a total of 900 tokens). The second corpus comprised /it#a/ vs. /i#ta/ and /Et#a/ vs. /E#ta/ sequences of various words structured as follows: "determinant + adjective (either "petit(e)" little, "maudit(e)" cursed or "parfait(e)" perfect) + noun", defined pairwise to be (quasi) minimal pairs, e.g., "une maudite amie" *'a cursed friend'* vs. "un maudit tamis" *'a cursed sieve'*. The phrases were embedded in the carrier sentence: "Il y a (…) ici" – '*There is (...) here'*, read by nine French female speakers (20 to 28 years) seven times (total: 1116 tokens). Bayesian mixed models were used for the statistical analyses. Separate models were run for each sequence type (/ap/, /aʁ/, /it/, and /Et/). We tested the effect of the position ("onset" for V_1.#C vs. "enchaînement" for V_1C#) on 4 acoustic measures: (i) the segmental acoustic duration of the consonant (closure + burst for plosives) and (ii) of V_1 , (iii) the F1 of V_1 (measured in the middle of the vowel), and (iv) the variability of the temporal lag between the acoustic middle of V_1 and acoustic middle of C, defined as the deviation of each V_1C lag from the median V_1C lag duration calculated for each speaker, condition and item.

Results and discussion. Our results differ according to the sequence. We found no difference for the /aʁ/ sequence in terms of segment duration or temporal variability. In Fig. 1 (left), the results are displayed for the segmental durations of V_1 and C. For /it/, /Et/ and /ap/ sequences, we found that in enchaînement the consonant duration is shorter than in onset position, but only for occlusive consonants. This result aligns with findings in other studies (see intro.). The V_1 is also shorter in enchaînement compared to onset in the sequences /ap/ and /it/, but is longer in the sequence /Et/. Thus, the results concerning the duration of the preceding vowel are inconclusive and exhibit different trends. Fig. 1 (right) displays the results for F1. F1 of the vowels /i/ and /E/ is higher in enchaînement than in the onset position, while the F1 of the vowel /a/ is lower in enchaînement than in the onset position. This indicates that these vowels can be considered as laxer in enchaînement position, as other studies have shown for high and mid-vowels (/ɛ/ in Fougeron 2007, /i/ and /o/ in L'Esperance 2015). The lower F1 for the vowel /a/ in enchaînement position can also be explained by a vowel laxing process in closed syllables (Storme, 2019), even if this is not phonological in French. Thus, it suggests that V_1 in

enchaînement retains its properties of being underlyingly in a close syllable. Moreover, the /VC/ lag variability shows that there is less variability in enchaînement position than in the onset whose segments are not tautosyllabic/tautolexical (see Fig. 2). This finding, which has not been investigated in other studies of resyllabification in French, suggests that there is probably a temporal coherence within the word that is not found at a word boundary even when *enchaînement* occurs. To conclude, we observe that the lexical unit remains preserved on the surface despite the phenomenon of *enchaînement* in French, primarily in terms of temporal stability (which is more stable within syllable/word) and secondarily in terms of preserved underlying syllabic structure.

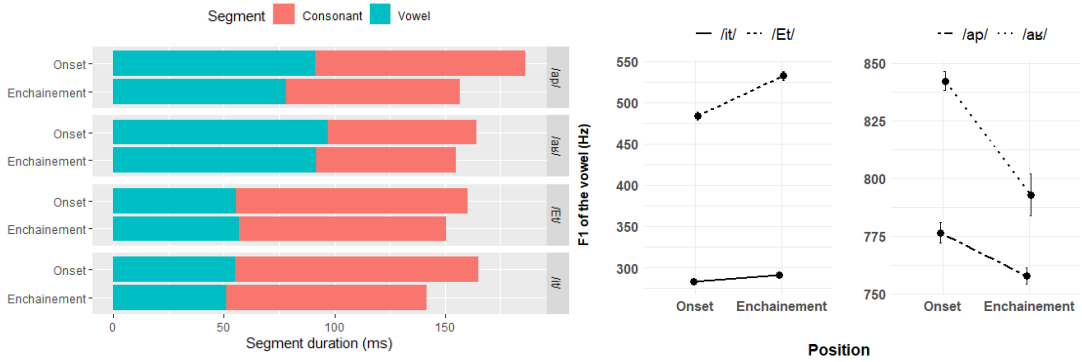


Figure 1: (left) *Segment duration (in ms) according to position and sequence. (right) F1 in preceding vowel (in Hz) according to the position and sequence.*

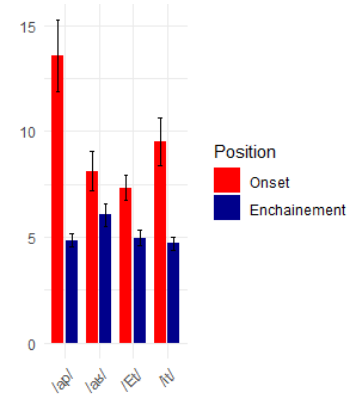


Figure 2: *Mean absolute deviation from the median of the VC lag according to the position and sequence.*

References

Bagou, O., Michel, V., & Laganaro, M. (2009). On the production of sandhi phenomena in French: Psycholinguistic and acoustic data. In *Proceedings of Interspeech-2009*, Brighton, U.K, 452–455.

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics*, *24*(2), 209-244.

Cho, T. (2001). Effects of morpheme boundaries on intergestural timing: Evidence from Korean. *Phonetica*, 58(3), 129-162.

Fougeron, C. (2007). Word boundaries and contrast neutralization in the case of enchaînement in French. *Papers in laboratory phonology IX: Change in phonology*, 609-642.

L'Esperance, M. J. (2015). *The Phonetics And Phonology Of Liaison Consonants In Montreal French*. [Master's Thesis, Cornell University]

Mücke, D. (2018). Dynamische Modellierung von Artikulation und prosodischer Struktur: Eine Einführung in die Artikulatorische Phonologie. *Language Science Press*.

Nam, H., & Saltzman, E. (2003, August). A competitive, coupled oscillator model of syllable structure. In *Proceedings of the 15th international congress of phonetic sciences*, 1, 2253-2256.

Rousselot, P.J. (1901). *Principes de phonétique expérimentale*. Paris : Welter.

Storme, B. (2019). Contrast enhancement as motivation for closed syllable laxing and open syllable tensing. *Phonology*, 36(2), 303-340.

Yersin-Besson, C. & Grosjean, F. (1996). L'effet de l'enchaînement sur la reconnaissance des mots dans la parole continue. *L'année Psychologique*, 96, 9-30.

Yoon, Y., Kim, S., & Cho, T. (2011). Stability of CV intergestural timing and coordination as a function of prosodic boundary and syllable structure in Korean. *17th International Congress of Phonetic Sciences*, Hong Kong, China.

Speech onset kinematics predict sentence level variability in adults who stutter

Torrey M Loucks[1], Daniel Aalto[2]

[1]Jacksonville University, Department of Communication Sciences and Disorders
[2]University of Alberta, Communication Sciences and Disorders
tloucks@ju.edu, aalto@ualberta.ca

Introduction. The onset of speech frequently conveys the most relevant linguistic and prosodic features of communicative intent. This transition into motion with its acoustic consequences could be more vulnerable to breakdown based on evidence that most stuttering disfluencies occur at speech initiation in people who stutter, Bloodstein & Bernstein Ratner (2008). Even if speech is initiated fluently, however, kinematic variability across a multisyllabic utterance is also higher in adults who stutter (AWS) compared to fluent adults (AWF), Loucks et al. (2022). We hypothesized the dynamics of the initial syllable of an utterance could predict the kinematic variability of a multisyllabic utterance. To test this hypothesis, we compared the kinematic characteristics of the first syllable of an utterance between AWS and AWF. Then we developed a regression model to test whether the first syllable movement characteristics determine the kinematic variability of a whole utterance.

Methods. Speech kinematic data were collected in 27 AWS (18-33 years; 20 males) and 26 AWF (19-16 years, 18 males), who were all native speakers of English. The AWS were diagnosed with stuttering as children and reported previous treatment for fluency. Stuttering severity was Mild for 9 participants, Moderate for 13 participants and Severe for 5 participants. The experiment was approved by the Ethics Review Board at the University of Alberta. The participants produced 15 repetitions of different multisyllabic real word and nonword phrases in randomized order. The initial syllable and whole phrase of the fluently produced stimulus sentence 'Buy Bobby a Puppy' are reported. The acoustic signal was acquired with a head-worn microphone at a 2" mouth-to-mic distance (sampling rate 44 kHz). Kinematic recordings of head, jaw and lip motion were acquired with the OptoTrak system (100 frames/sec). The onset of lower lip (LL) opening for 'buy' and the peak closing of the final 'p' in puppy were marked in the kinematic amplitude record of 10 tokens for each participant (inferior-superior dimension). The LL kinematic vectors of the first syllable and whole phrase were then normalized in the spatial and temporal domains (1000 points) providing separate estimates of the Spatiotemporal Index (STI) for the syllable and phrase, as in Smith et al. (1995). In a separate analysis, the following kinematic points of the LL for the first syllable 'buy' in the same dimension were labelled with automated algorithms: 1) peak opening amplitude (mm), 2) peak opening velocity (mm/sec), and, 3) opening time (sec).

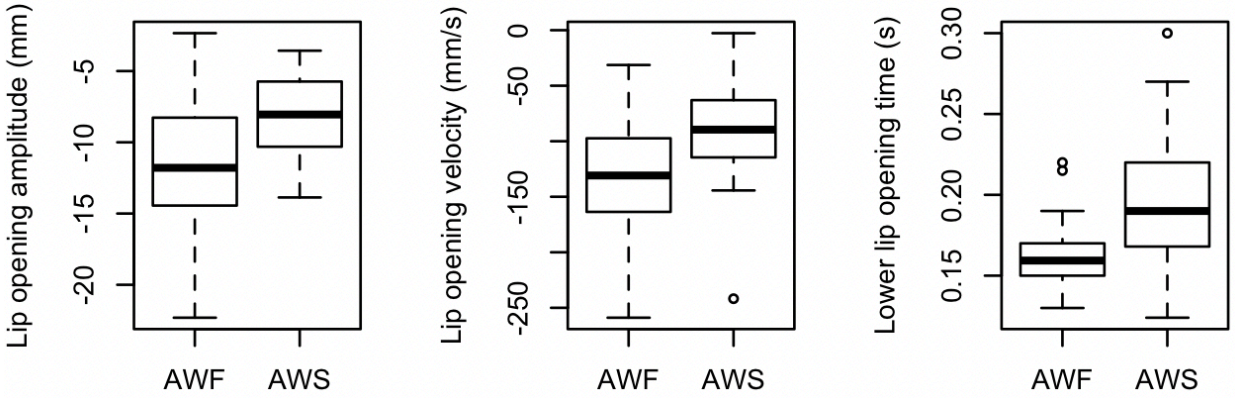


Figure 1. Boxplots for lip opening amplitude, velocity, and opening time for adults who are fluent and who stutter.

Results. The LL motions of the AWS for the first syllable had significantly smaller amplitude (t=3.2, p=.0024), lower peak velocity (t=3.2,p=.0021), and opening time (t=4.1,p=.0002) than AWF (Figure 1). The lower lip STI of the AWS was significantly higher for both the initial syllable (t=-3.7, p=.0009) and the whole phrase compared to AWF (t=5.7, p=2.5e-6). Significant correlations between each of the predictors and the whole phrase STI were found for the AWS ($p<.01$ for each correlation), but these correlations were not significant for AWF. The initial syllable movement characteristics were fitted in a regression model to predict whole phrase variability for AWF and AWS separately (dependent variable: STI; independent variables: amplitude, velocity, duration). In the AWF, initial syllable

characteristics did not predict phrase STI scores (Adjusted R-squared: -.01). In contrast, there was good prediction of the phrase STI for the AWS. A model with initial syllable LL opening amplitude and LL opening time predicted phrase STI (Adjusted R-squared: .53) with an intercept of 15.3 (SE 4.5, t=3.4, p=.002), and slopes of 1.1 and 37.9 for LL opening amplitude and LL opening time, respectively (SE 30, t=3.8, p=.001; SE 9.5, t=4.0, p=.0005). The model did not improve (Akaike information criterion) when initial syllable STI and/or peak velocity were added as dependent variables. Residual plots do not suggest violations of model assumptions or undue impact of single observations.

Discussion. In this study, the speech initiation pattern of AWS involved slower and smaller amplitude LL movements that diverged into more variable kinematic trajectories across the whole phrase compared to AWF. Further, these altered speech onset patterns predicted the higher whole phrase variability of the AWS. Van Lieshout et al. (1996) also identified aberrations in the speech onset of AWS involving delays and increased amplitude of LL surface EMG relative to AWF. Smith & Kleinow (2000) found that AWS tended to have lower amplitude and lower velocity of LL motion using the same phrase, but there were no statistical differences compared to AWF and the kinematic variables did not predict STI. Our current study differed from these earlier studies in focusing on the initial syllable and having substantially more participants. Recent neurological investigations point to atypical brain activity preceding the onset of both fluent speech and stuttering disfluencies in AWS (Sengupta et al. 2017; Mersov et al. 2016), in addition to the studies demonstrating that most stuttering occurs at speech onset (Bloodstein & Bernstein Ratner 2008). AWS may have adopted a more cautious speech initiation pattern to avoid disfluencies or due to previous therapy involving speech onset strategies. Alternately, the higher variability and altered kinematics reported here could reveal aberrations in speech planning that are manifested at movement onset for AWS. The susceptibility to kinematic aberrations at onset could then cascade into higher articulatory variability for the whole phrase. The AWF speakers did not show a relationship between speech onset and whole phrase variability despite substantial inter-speaker variation. The larger amplitude and rapid motions of AWF could reflect robust speech planning and motor programming that is maintained across the phrase. This study provides novel findings that kinematic point measures of speech initiation are possibly related to global variability in stuttering. Future research should be broadened to investigate: 1) additional utterances, 2) more repetitions, 3) LL EMG, 4) other articulation patterns (e.g., pre/post therapy), 5) rate variations, and 6) intonation variations.

References

1. Bloodstein, O., & Bernstein Ratner, N. (2008). *A handbook on stuttering*. Clifton Park, NY: Delmar
2. Loucks, T. M., Aalto, D., Lomheim, H., & Pelczarski, K. (2022). Speech kinematic variability in Adults who Stutter is influenced by treatment and speaking style. *Journal of Communication Disorders. 96*, 106194. https://doi.org/10.1016/j.jcomdis.2022.106194
3. Smith, A., Goffman, L., Zelaznik, H. N., Ying, G., & McGillem, C. (1995). Spatiotemporal stability and patterning of speech movement sequences. *Experimental Brain Research, 104,* 493–501
4. Van Lieshout, P., Hulstijn, W., & Peters, H. F. (1996). From Planning to Articulation in Speech Production: What Differentiates a Person Who Stutters From a Person Who Does Not Stutter? *Journal of Speech, Language, and Hearing Research*, 39(3), 546–564
5. Smith, A. & Kleinow, J. (2000). Kinematic correlates of speaking rate changes in stuttering and normally fluent adults. *Journal of Speech, Language and Hearing Research*, 43(2), 521.
6. Sengupta, R., Shah, S., Loucks, T. M., Pelczarski, K., Yaruss, J. S., Gore, K. & Nasir, S. M. (2017). Cortical dynamics of disfluency in adults who stutter. *Physiological Reports*, 5(9), e13194-13205, doi: 10.4814/phy2.13194
7. Mersov, A. M., Jobst, C., Cheyne, D. O., & De Nil, L. (2016). Sensorimotor Oscillations Prior to Speech Onset Reflect Altered Motor Networks in Adults Who Stutter. *Frontiers in Human Neuroscience*, *10*, 443. https://doi.org/10.3389/fnhum.2016.00443

Acoustic characteristics of narrative elements in infant-directed speech

Anna Kohári[1], Katalin Mády[1], Uwe D. Reichel[1], Veronika Harmati-Pap[1], Bence Kas[1,2], Sarolta Murányi[3]

[1]HUN-REN Hungarian Research Centre for Linguistics
[2]Eötvös Loránd University, MTA-ELTE Language-Learning Disorders Research Group
[3]ELTE Department of Applied Linguistics and Phonetics

kohari.anna|mady.katalin|uwe.reichel|harmati.pap.veronika|
kas.bence@nytud.hun-ren.hu, muranyi.sarolta@btk.elte.hu

Introduction. Speakers aim to adapt their speech to the needs and mental abilities of their audience, especially when addressing an infant. Among many other factors, the linguistic environment, including the characteristics of infant-directed speech (IDS), can influence the rate of a child's language development (Saint-Georges *et al.* 2013). However, little is known about the diverse speech adaptation strategies parents use in talking to their children. IDS exhibits several acoustic features, such as higher pitch, slower speech rate, and often larger vowel space compared to adult-directed speech in several languages (Hilton *et al.* 2022). These characteristics of IDS were also found in the IDS of native Hungarian mothers (Gergely *et al.* 2017; Kohári & Mády 2023). Usually, acoustic features are examined separately, without comparing them to linguistic differences. Moreover, communication with children often takes the form of telling stories based on pictures, and the macrostructural elements of storytelling in IDS, as well as the functions of speech units in discourse, is rarely investigated. It has been observed that parents frequently use various conversational elements, including asking questions during storytelling (cf. Disbray 2008). In this research, we investigate the strategies speakers apply during storytelling to their children and how these strategies relate to the acoustic features of speech.

Methods. We compared the recordings of 90 native Hungarian mothers narrating a story to their 6-month-old infants (IDS) and to an experimenter (adult-directed speech, ADS). Mothers were instructed to tell the story in a semi-spontaneous manner, based on the images in a colourful story book but to incorporate 17 prescribed sentences word by word. Recordings were conducted using a Beyerdynamic TG H74c supercardioid head-mounted condenser microphone at the baby lab of the Research Centre for Natural Sciences of the Hungarian Academy of Sciences. For the acoustic analysis, only the 'mandatory' scripted sentences were evaluated for the sake of comparability between the two registers. We measured the following variables: speech rate, articulation rate, vowel space dispersion (similarly to Karlsson & Doorn 2012), f0 median, minimum, maximum, and range in semitones using the CoPaSul tool (Reichel 2016). We calculated the arithmetic means for each mother separately for the two conditions (IDS and ADS), and the difference between their IDS- and ADS-means. For the analysis of the story grammar, we counted the communication units (C-units, Cebioğlu *et al.* 2022) formulated based on the images and further analyzed the macrostructure of the narratives (Wigglesworth & Stavans 2001; McArthur *et al.* 2005). On the one hand, we measured what percentage of the C-units involve so-called narrative units that contain information crucial to the story but not present in the scripted sentences (specifically in the story: bird, boat, hat, lion's mane, walnut, color names). On the other hand, we determined what percentage of C-units contain so-called conversational elements not directly related to the story, but aiming to engage the child in the storytelling (e.g., *Look!*; *See, here are the pixies!*). Based on the scaled values of the three storytelling metrics, we performed cluster analysis (Partitioning Around Medoids PAM) using the package cluster (Maechler *et al.* 2023) in R software, and compared the resulting groups with the acoustic data quantifying the speakers' IDS production and the differences between their speech in the two registers (Wilcoxon test).

Results. In order to determine the storytelling strategy of each speaker, the three narrative characteristics (number of C-units, percentage of narrative units, percentage of conversational elements) were scaled and clustered based on the difference of the variables between the two registers, using the PAM clustering algorithm. The optimal number of clusters that maximized the mean silhouette score was found to be 2. The first cluster included speakers who were inclined to use a higher percentage of conversational elements in IDS than in ADS. The speakers of this cluster typically told much longer stories in IDS than in ADS. The second cluster comprised speakers who typically told shorter stories, used fewer conversational elements and proportionally more narrative units in IDS than in ADS. We compared these two clusters with the acoustic data on these speakers' IDS and on their ID-AD differences to understand how speaker strategies in storytelling and speech acoustics relate to each other (Figure 1). The results showed that those who used conversational elements more frequently in IDS (Cluster 1) had a larger difference in f0 median and f0 maximum between the two registers than the speakers in the other cluster (f0 maximum: $W = 1225$, $p = 0.006$; Cliff's Delta = 0.351, f0 median: $W = 1228$, $p = 0.006$, Cliff's Delta = 0.103). Therefore, the more frequent use of conversational elements in IDS was associated with speakers who separate IDS from ADS in a more distinct manner in terms of the f0-related variables. Among the

above variables only the ID-AD differences of the metrics exhibited significant differences between the clusters; the measured values themselves in the IDS condition were similar in both clusters ($p > 0.05$). Furthermore, in the cluster of speakers who used conversational elements more frequently (Cluster 1), the vowel space measured in IDS was larger than in those who used this strategy less often ($W = 873$, $p = 0.024$; Cliff's Delta = 0.595). Speech rate, articulation rate, as well as the f0 minimum and f0 range metrics did not show any correlation with the clusters of narrative elements. Regardless of the clusters, we also examined how the speakers differentiate their speech when talking to infants or to adults. We found that almost every acoustic and story grammar variable exhibited differences between the two registers, except for f0 range and the percentage of narrative units.

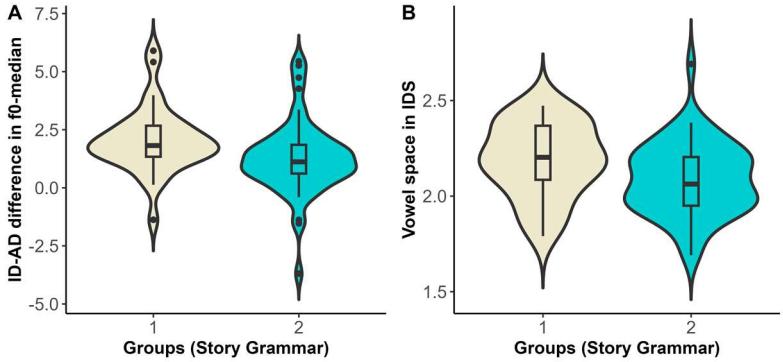


Figure 1: *ID-AD difference in f0 median (A) and vowel space in IDS (B) for the two clusters. Cluster 1 shows, e.g., that conversational elements appear even more frequently in IDS than in ADS, compared to Cluster 2.*

Discussion. Our results revealed a connection between the macrostructure of the stories told in infant-directed semi-spontaneous storytelling and the acoustic properties of speech. Speakers who tended to narrate their stories longer with more conversational elements differentiated infant-directed speech from adult-directed speech more, based on the f0-related variables. This phenomenon may contribute to capturing and maintaining the child's attention (Saint-Georges *et al.* 2013). The same speakers typically used a larger vowel space in their infant-directed speech than those who applied fewer conversational elements and told shorter stories, possibly supporting the vocabulary development of their children (cf. Kalashnikova & Burnham, 2018). The results draw attention to the phenomenon that speakers not only modify the acoustic properties but also the narrative elements of their speech directed to their children, using different strategies to exploit these characteristics.

Acknowledgements. This study was funded by the National Research, Development and Innovation Office, grants K 115385, PD 134775 and K 124477. Anna Kohári and Sarolta Murányi are the lead authors of the study.

References

Cebioğlu, S., Marin, K. A., Broesch, T. (2022). Variation in Caregivers' References to their Toddlers: Child-directed Speech in Vanuatu and Canada. *Child Development, 93*(6), e622-e638. DOI: 10.1111/cdev.13833

Disbray, S. (2008). Storytelling styles: a study of adult–child interactions in narrations of a picture book in Tennant Creek. In J. Simpson & G. Wigglesworth (Eds.), Children's language and multilingualism: Indigenous language use at home and school. New York: Continuum. 56-78.

Gergely, A., Faragó, T., Galambos, Á., & Topál, J. (2017). Differential effects of speech situations on mothers' and fathers' infant-directed and dog-directed speech: An acoustic analysis. *Scientific Reports*, 7(1), 13739.

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., Simson, J., ... & Mehr, S. A. (2022). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour, 6*(11), 1545-1556.

Kalashnikova, M., & Burnham, D. (2018). Infant-directed speech from seven to nineteen months has similar acoustic properties but different functions. *Journal of Child Language*, *45*(5), 1035-1053.

Karlsson, F., & Doorn, J. V. (2012). Vowel formant dispersion as a measure of articulation proficiency. *The Journal of the Acoustical Society of America, 132*(4), 2633-2641.

Kohári, A., & Mády, K. (2023). A longitudinal study of pauses, interpausal units and clauses in infant-directed speech. In R. Skarnitzl & J. Volín (Eds.), Proceedings of 20th International Congress of Phonetic Sciences (ICPhS). Praha: Guarant International. 2369-2373.

Maechler M., Rousseeuw P., Struyf A., Hubert M., Hornik K. (2023). *cluster: Cluster analysis basics and extensions*. R package version 2.1.5. https://CRAN.R-project.org/package=cluster.

McArthur, D., Adamson, L. B., Deckner D. F. (2005). As Stories Become Familiar: Mother-Child Conversations During Shared Reading. *Merrill-Palmer Quarterly, 51*(4). 389-411.

Reichel, U. D. (2016). *CoPaSul manual. Contour based, parametric, and superpositional intonation stylization*. arXiv:1612.04765.

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., Laznik, M.-C., & Cohen, D. (2013). Motherese in interaction: at the cross-road of emotion and cognition? (A systematic review). *PloS One, 8*(10), e78103.

Wigglesworth, G., Stavans, A. (2001). A Crosscultural Investigation of Australian and Israeli Parents' Narrative Interactions with Their Children. In: Nelson, K. E., Aksu-Koç, A., Johnson, C. E. (eds.). Children's language: Developing narrative and discourse competence. New Jersey: Lawrence Erlbaum Associates. 73-91.

Lingual and epilaryngeal articulation of vowels in Mundabli

Author Full Name[1], Co-author Full Name[2]

[1]*Affiliation*
[2]*Affiliation*
Author 1 email address, Author 2 email address

Introduction. Mundabli is a Yemne-Kimbi language spoken in the Lower Fungom region of northwestern Cameroon (Good et al. 2011). Mundabli's vowel system is rich in contrasts based on lower vocal tract activity, featuring ten plain monophthongs /i ɪ e ɛ ɨ a u ʊ o ɔ/ and six monophthongs which have been described as pharyngealized /iˤ eˤ ɨˤ aˤ uˤ oˤ/ (Voll 2017; Faytak et al. 2023). The pharyngealized vowels have developed recently in the language's history from vowels followed by coda **k*; the most closely related neighboring language, Mufu, has /k/ or /ʔ/ in numerous cognate lexical items, realized as [k] [ʔ] depending on the vowel (i.e. Mufu [kjōk], Mundabli [tsōˤ] 'banana'; Mufu [bàʔ], Mundabli [bàˤ] 'scar'; Mufu [dàk], Mundabli [dèˤ] 'place'). The contrast between "tense" /i e u o/ and "lax" /ɪ ɛ ʊ ɔ/ may be better characterized as a distinction of tongue root advancement or retraction, common in the broader West African region (Kirkham and Nance 2017). It is unclear how Mundabli speakers would organize articulation to accommodate a three-way distinction in lower vocal tract activity since such a situation is vanishingly rare in the world's languages. As such, this exploratory study aims to clarify the acoustic differences between Mundabli's pharyngealized and plain vowel sets, and the lingual and epilaryngeal articulatory basis of these distinctions.

Methods. Time-aligned acoustic and ultrasound data were collected from 15 speakers in 2022 and 2023 in Douala, Cameroon. Stimuli were open-syllable words containing all 16 Mundabli vowels (two types per vowel) read in a frame sentence verbally prompted by the first author. Voice quality measures (H1*-H2*, CPP) and formant frequencies (F1-F3) were extracted using PraatSauce at nine evenly spaced time points across vowels' durations. All measures were z-scored and outliers removed; the final acoustic data set contains 42,539 timepoint measurements (roughly 4,726 vowel tokens). GAMs were carried out on the acoustic data to assess the evolution of each feature over the vowels' durations. The co-collected ultrasound data was also analyzed for two speakers. Tongue surface splines and hyoid position were extracted using DeepLabCut as implemented in Articulate Assistant Advanced v220.2. Splines and hyoid positions were extracted at vowel midpoint and endpoint based on annotations of the acoustic data.

Results. As a group, pharyngealized vowels exhibit elevated CPP and lowered H1-H2* relative to their plain counterparts, suggesting tense or creaky phonation (Figure 1A). They also exhibit raised F1, raised F2, and lowered F3 (Figure 1B) relative to plain vowels. Most measures remain distinct over the entire duration of the vowel, with the exception of H1-H2*.

The ultrasound data demonstrate that pharyngealized vowels exhibit a raised tongue dorsum and a lower pharyngeal constriction (1C) relative to their plain counterparts; "double-bunching" is observed for all of the pharyngealized vowels to some extent, with a concave tongue dorsum shape often evident. This tongue shape holds across the whole duration of the vowel, but strengthens slightly towards the end of the segment's duration. The "tense" vowels /i e u o/ are realized with slightly fronted tongue root compared to the "lax" vowels /ɪ ɛ ʊ ɔ/; differences in dorsum height are not consistently observed. Hyoid involvement with pharyngealization is minimal; the variation in hyoid position observed is almost entirely related to f0 variation.

Discussion. The Mundabli pharyngealized vowels exhibit a constriction in the lower pharynx. The presence of tense or creaky voice on pharyngealized vowels, which is not otherwise contrastive in Mundabli, suggests that constriction of the epilaryngeal tube plays some role as well. This is distinct from other lower vocal tract phenomena often called "pharyngeal", such as emphasis (Al-Tamimi 2017) or uvularization (Evans et al. 2016). Lower pharyngeal and/or epilaryngeal constriction is clearly suggested by both the F2-raising effect seen in the pharyngealized vowels (as opposed to

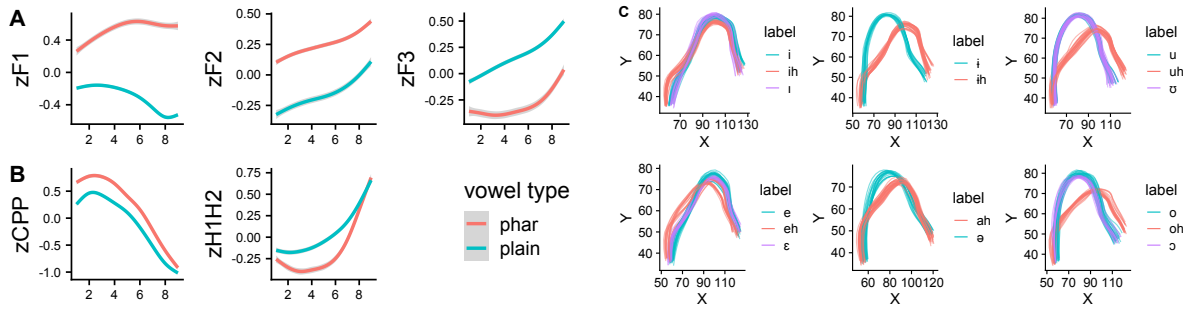


Figure 1: *GAM smooths of formant frequencies (A) and voice quality (B) for all 15 speakers, and raw traces for pharyngealized and plain vowels (C) for one representative speaker. The tongue tip is oriented to the right.*

F2-lowering for uvularization) and the characteristic double-bunching also observed in languages with lower pharyngealization Catford 1983; Arkhipov et al. 2019. Pharyngealization holds across the entire vowel, suggesting a total loss of any historical coda consonant; further study of Mundabli and its neighboring languages may improve our understanding of of coda consonant vocalization and loss generally.

References.

Arkhipov, A, M Daniel, O Belyaev, G Moroz, and JH Esling (2019). "A reinterpretation of lower-vocal-tract articulations in Caucasian languages". In: *Proceedings of the 19th International Congress of the Phonetic Sciences*, pp. 5–9.

Catford, John C (1983). "Pharyngeal and laryngeal sounds in Caucasian languages". In: *Vocal fold physiology: Contemporary research and clinical issues*, pp. 344–350.

Evans, Jonathan P, Jackson T-S Sun, Chenhao Chiu, and Michelle Liou (2016). "Uvular approximation as an articulatory vowel feature". In: *Journal of the International Phonetic Association* 46.1, pp. 1–31.

Faytak, Matthew, Bowei Shao, Angèle Douanla Taffre, and Nelson C Tschonghongei (2023). "Frication and formant frequencies in the Mundabli high vowels". In: *20th International Congress of Phonetic Sciences*.

Good, Jeff, Jesse Lovegren, Jean Patrick Mve, Carine Nganguep Tchiemouo, Rebecca Voll, and Pierpaolo Dicarlo (2011). "The languages of the Lower Fungom region of Cameroon: Grammatical overview". In: *Africana Linguistica* 17.1, pp. 101–164.

Kirkham, Sam and Claire Nance (2017). "An acoustic-articulatory study of bilingual vowel production: Advanced tongue root vowels in Twi and tense/lax vowels in Ghanaian English". In: *Journal of Phonetics* 62, pp. 65–81.

Al-Tamimi, Jalal (2017). "Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: Implications for formal representations". In: *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8.

Voll, Rebecca M (2017). *A grammar of Mundabli, a Southern Bantoid (Yemne-Kimbi) language of Cameroon*. Netherlands Graduate School of Linguistics.

Decoding orofacial signals beyond sight: A study of expressive faces and whispered voices in German

A simple look at an interaction between a pair of speakers can reveal the fact that human communication is more than the mere exchange of words and involves a multimodal system in which gestures play an integral role. In this domain, one of the intriguing questions is on the nature of the relationship between gesture and speech. Closely related to this, there are two influential hypotheses. One possible conjecture is that there is a trade-off relation between gesture and speech in terms of the communicative load [1], [2], [3], and [4]. Another alternative account is hand-in-hand hypothesis which views the relation between gestures and speech in parallel rather than compensatory [5], [6]. These two hypotheses largely depend on type of gesture as well as the communicative settings and constraints [7], [8]. In this study, we focus on measuring the orofacial expressions including eyebrow movements, eye opening, and lip aperture in two different prosodic conditions, i.e., polar questions with rising intonation vs. statements with falling or flat intonation. The varying intonation can enable us to find out whether and to what extent speech with varying prosody interacts with the oro-facial expressions. Taking "whispered/semi-whispered speech" and "(in)visibility" of speakers as two communicative and cognitive difficulties into account, we aim to investigate what happens to speech and gesture in situations where speakers whisper and do not see each other.

To this end, we ran an experiment in which 20 native speakers of German were simultaneously audio and video recorded while producing 20 pairs of statement and questions. The content of questions vs. statements was identical with the only difference in the punctuation mark. Each sentence was composed of 4 content words and the target word which was the focus of our study, was at the sentence's final position. All the target words were bisyllabic with the stress falling on the initial syllable. The stressed syllables had CVC structure containing one of the bilabial stops /p/, /b/, /m/ followed by an unrounded vowel of /e/, /a/, /i/. The experiment took part in the interaction between a confederate and a participant. The confederate who was the same speaker during the whole experiment, generated either a question or a statement and the participants were supposed to respond the question by converting it into a statement or ask a question in response to the statement by altering their intonations (see appendix). The experiment consisted of four stimulus blocks linking two conditions, i.e., speech mode [*normal, semi-whispered, and whispered speech*] and visibility [*visible vs invisible mode*]. The orofacial expressions were detected and measured using Openface.2 facial landmark detector [9] by which each video was iterated, and 68 landmarks were mapped into the key regions of the face.

Based on the results of linear mixed effect models, the interactions between left/right eyebrows and visibility condition were significant (right eyebrow $t= 2.633$, $p<.01$, left eyebrow $t=3.903$, $p<.001$). The three-way interaction between *Speech Mode*(In)visibility*Sentence Type* for the left eyebrow ($t=1.886$, $p=.05$) was at the level of statistical tendency indicating that speakers raise their eyebrows the highest when they produce questions in whispered speech and when they are invisible. For the lip aperture, the significant interaction between *Speech Mode* Sentence Type* ($t= 2.280$, $p<.05$) was reflected in a larger difference between whispered, semi-whispered and normal speech when questions are produced. The results also revealed a significant effect between *Speech Mode*(In)visibility*Sentence Type* for both eyes ($t= -2.837$, $p < 0.01$ for the left eye, and $t= -2.672$, $p < .01$ for the right eye) with the largest effect for the questions produced in semi-whispered mode of speech and when speakers do not see each other.

Overall, the results reveal more pronounced oro-facial expressions in communicatively marked situations, i.e. when speakers do not see each other and when they whisper. These findings replicate our results found for a typologically different language, i.e. Farsi.

Appendix

Stimuli sample:

Question condition

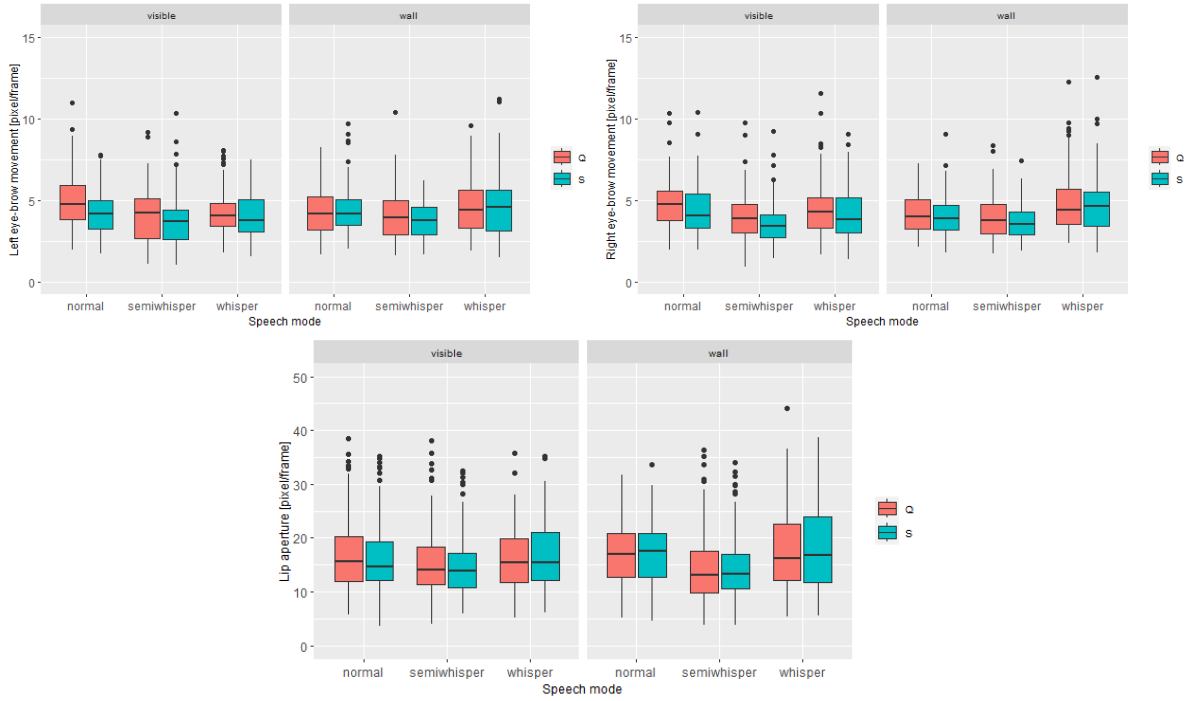
Confederate: Er hat eine Bitte.
"He has a request."

Informant: Er hat eine Bitte?
"He has a request?"

Statement condition

Confederate: Er hat eine Bitte?
"He has a request?"

Informant: Er hat eine Bitte.
"He has a request."



References

[1]. Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science, 15*(6), 415-419.

[2]. Melinger, A., & Levelt, W. J. (2004). Gesture and the communicative intention of the speaker. *Gesture, 4*(2), 119-141.

[3]. De Ruiter, J. P. (2006). Can gesticulation help aphasic people speak, or rather, communicate?. *Advances in Speech Language Pathology, 8*(2), 124-127.

[4]. Van der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes, 44*(3), 145-174.

[5]. Goldin-Meadow, S. (2009). How gesture promotes learning throughout childhood. *Child Development Perspectives, 3*(2), 106-111.

[6]. So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science, 33*(1), 115-125.

[7]. Bavelas, J. (1994). Gestures as part of speech: Methodological implications. *Research in Language and Social Interaction, 27*, 201–221.

[8]. Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language, 58*, 495–520.

[9]. Baltrušaitis, T., Zadeh, A., Lim, Y. Ch., Morency L.-Ph. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*. https://par.nsf.gov/servlets/purl/10099460

Linking levels of prosodic structure to modulatory activity in speech production

Leonardo Lancia[1], Jinyu Li [2], Caterina Petrone[1], Louis Goldstein[3]

[1]Laboratorie Parole et Langage, Aix-Marseille Université / CNRS
[2]Laboratoire de Phonétique et Phonologie, Université Sorbonne Nouvelle / CNRS
[3]Department of Linguistics, University of Southern California

Leonardo.lancia@cnrs.fr, jinyu.li@sorbonne-nouvelle.fr, caterina.petrone@univ-amu.fr, louisgol@usc.edu

Introduction. The production of speech units of different levels (e.g. gestures, syllables, accents etc.) introduces regular fluctuations in the properties of the acoustic waveform. For example, from acoustic amplitude or from sonority envelopes we can extract oscillations depending on syllables production (Wang and Narayanan, 2007) or on word level prominence; while from measures of spectral change, we can extract oscillations depending on vocal tract movements implementing articulatory gestures (Goldstein, 2019). In biological systems, cyclic patterns of activity often have a regulatory function structuring the joint behavior of different organs or organisms over time. This work describes a two-steps analysis to reveal such an organization in speech signals. In the first step, we decompose the speech signals into oscillations at time scales of several linguistic units (from segments to accented syllables); in the second step we test how well these oscillations permit predicting the locations of the frontiers between linguistic units individuated via manual segmentation. Our finding show: 1) that oscillatory patterns of activity extracted from the speech signal are related to the organization of timing at all tested levels of the prosodic structure; 2) that oscillations extracted from different features permit predicting the behavior of units at different levels in a language specific fashion; and 3) that the oscillatory patterns reflect the correlations across dimensions of speech production.

Method. We applied our approach to two corpora of French and English readings manually segmented (8 and 10 speakers respectively extracted from Chanclu et al., 2020 and from Grabe et al., 2001). Oscillatory components were extracted from the following features: 1) the spectral flow (as estimated from the total changes of spectral energy following Goldstein 2019); 2) the acoustic energy; 3) the f0 (interpolated via cubic splines in unvoiced regions). From the continuous speech streams, we extracted 131 chunks not shorter than four seconds and separated by unvoiced intervals longer than 200 msec. After extracting the acoustic features from each uninterrupted recording, the following steps were applied to each chunk separately. To extract the oscillatory components, we applied FIR band-pass filters centered at the rates of occurrence of sub-syllabic units, syllabic units and of perceived accents, as inferred from the manual segmentation of the relevant units. The supra-syllabic units correspond to Accentual Phrases in French, bounded at the right by the perceptual prominent syllable, and Phonological Phrases in English bounded on the right by a perceptual prominent syllable often corresponding to the nuclear phrase accent. Some boundaries delimit also Intonational Phrases, but they were not distinguished. The oscillatory signals obtained were submitted to amplitude demodulation and to the Hilbert transform in order to extract the instantaneous phase (growing from 0 to 2π in each cycle, see Lancia, 2023 for details). A phase signal was also extracted from the boundaries of the speech units at each structural level considered. The phase signal grows linearly from 0 to 2π during the time interval corresponding to the production of each unit and it remains at zero during pauses (see **Figure 1**). Note that by assigning phase values to each speech unit, we map it onto an ideal cycle of activity.

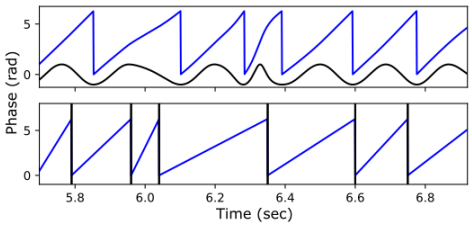


Figure 1: *Phase extraction from the demodulated oscillatory signal in the topmost panel (black curve) and from the boundaries in the bottommost panel (black vertical lines). Phase is in blue in both panels.*

To estimate the stability of the temporal alignment between each oscillatory component and the locations of the boundaries at each structural level, we computed the Phase Locking Value (PLV, see Lancia et al., 2023 for details), which is inversely related to the variability of the difference between two phase signals in each speech chunk. PLV grows as the phase of the oscillatory signal permits predicting with increasing precision the locations of the boundaries. In practice, PLVs obtained in this way are affected by the choice of the filter frequencies, which is in turn informed by a-priory knowledge of boundary locations. This dependency introduces a degree of circularity (the filter frequencies depend

on the number of boundaries whose predictability is estimated through the PLV) and a number of confounding factors (e.g.: a more variable unit rate will provide a less precise estimate of the instantaneous oscillation frequency). In order to make the computed PLVs independent of these effects, each observed PLV is normalized (nPLV=PLV/rPLV). rPLV is the PLV expected between the observed oscillation and randomized sequences of boundaries obtained by shuffling the durations between consecutive observed boundaries in the analyzed chunk. rPLV is computed by averaging the results obtained with 30 such randomized sequences of boundaries. Higher than chance coordination is observed when nPLV>1.

Results and discussion. Different columns of **Figure 2** contain the nPLVs obtained by extracting oscillatory signals from different languages and acoustic features. The horizontal axis differentiates between different unit kinds (segments, syllables, chunks delimited by perceptually prominent syllables), colors indicate the time scale of the modulatory signal considered (with blue, red and green respectively associated to the sub-syllabic, syllabic and supra-syllabic time scales). A filled circle indicates that the nPLVs are not significantly larger than one (which would indicate lack of significant temporal coordination) according to one sided t-tests corrected via the false discovery rate criterion. A cross indicates the presence of significant coordination according to the same test.
Not surprisingly, spectral flow signals are more adapted to capture sub-syllabic oscillations; while acoustic energy is better suited to extract syllabic oscillations. Only in French, f0 oscillations at the accentual level are the best predictor of boundaries at that level. This finding suggests that the relation between the slow oscillations of the f0 signal and structural prosodic levels differs across the two languages. It also raises the possibility that the time scale adopted to characterize the supra-syllabic oscillations based on perceived pitch accents may not be appropriate for English. A language specific definition of the supra-syllabic time-scale, which for English could be based on stress location, is likely to be more suitable and it is currently being investigated.
Units whose boundaries are better predicted through a given signal, display significant relations with several oscillatory scales. This can be explained by the expected mutual dependency between oscillations at different time scales (Leong et al., 2014; Lancia et al., 2018). All acoustic features (except spectral flow in French) permit predicting several levels of annotation via oscillations at different time scales. This result most likely depends on known correlations between dimensions of activity (e.g. f0 and energy), however it also shows that rhythmic regularities (here the oscillations at different time scales) reflect the coordination of speech production activity across different dimensions (here the different acoustic features). In other words, mutual influences between different dimensions of speech production potentially serving different linguistic purposes, are integrated in a language specific and globally coherent hierarchy of time scales organizing the timing of speech production at several levels of the prosodic structure.

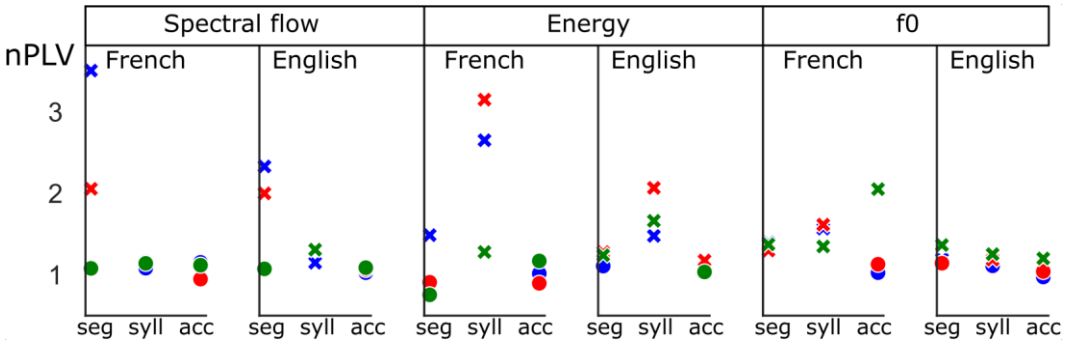


Figure 2: *Mean Phase Locking Values between different oscillatory scales (color coded) and units at different structural levels (distributed on the x axis) for each acoustic feature and language (in different columns).*

References

Chanclu, A., Georgeton, L., Fredouille, C., & Bonastre, J.-F. (2020) PTSVOX: une base de données pour la comparaison de voix dans le cadre judiciaire, in 6e conférence conjointe Journées d'Études sur la Parole, (pp. 73–81).

Goldstein. (2019). The role of temporal modulation in sensorimotor interaction. *Frontiers in Psychology*, *10*, 2608.

Grabe, E., Post, B. & Nolan, F. (2001). Modelling intonational Variation in English. The IViE system. In Puppel, S. and Demenko, G. (eds). *Proceedings of Prosody 2000.* Adam Mickiewitz University, Poznan, Poland.

Lancia, L., Krasovitsky, G., & Stuntebeck, F. (2019). Coordinative patterns underlying cross-linguistic rhythmic differences. Journal of Phonetics, 72, 66-80.Lancia, L. (2023). Instantaneous phase of rhythmic behaviour under volitional control. *bioRxiv*, 2023-11.

Lancia, L. (2023). Instantaneous phase of rhythmic behaviour under volitional control. *bioRxiv*, 2023-11.

Lancia, L., Li, J., & Fougeron, C. (2023). How speech rate, syllabic complexity and diversity affect the emergence of speech rhythm in speeded syllable repetition. In *Proceedings of ICPHS 2023.*

Leong, V., & Goswami, U. (2014). Assessment of rhythmic entrainment at multiple timescales in dyslexia: Evidence for disruption to syllable timing. *Hearing research*, *308*, 141-161.

Wang, D., & Narayanan, S. S. (2007). Robust speech rate estimation for spontaneous speech. IEEE Transactions on Audio, Speech, and Language Processing, 15(8), 2190-2201.

Sibilant contrast production by bilingual speakers of Quanzhou Southern Min and Mandarin

Caihong Weng[1], Ioana Chitoran[1], Alexander Martin[2]

[1]*Université Paris Cité*

[2]*University of Groningen*

caihong.weng@etu.u-paris.fr, ioana.chitoran@u-paris.fr, alexander.martin@rug.nl

Introduction. A merger of the Mandarin sibilant fricative contrast [s]~[ʂ] has been observed in Mandarin spoken by bilingual L1 Southern Min speakers, a phenomenon commonly characterized as "deretroflexion". Previous work has suggested that language contact with Southern Min, which lacks the retroflex phone, is the primary source of the merger (Kubler 1985). Yet, variability in these bilingual speakers' productions remains to be explained. Recent research suggests that the variability in the merger of this sibilant contrast is best captured by considering social (e.g., age, gender, and individual-level exposure to and use of Mandarin) in addition to linguistic factors (Chang and Shih 2015; Chuang and Fon 2010; Lee-Kim and Chou 2022). The present paper explores variation in the production of the Mandarin [s]~[ʂ] contrast among a sample of bilingual speakers of Quanzhou Southern Min (QSM) [L1] and Mandarin [L2] and examines different linguistic and social factors.

Method. 60 bilingual speakers of QSM and Mandarin (28 men, 32 women) were recruited in Quanzhou, China, divided into three age ranges between 18 and 60 (18–30: 27 participants, 31–40: 18 participants, 41–55: 15 participants). Each participant took part in a sentence reading task with target words embedded into carrier sentences. Targets were all real Mandarin words of the form CVCV (2 fricatives $\times$ 2 vowel contexts ([a] vs. [u]) $\times$ 3 examples) and realized with a high level tone (tone 1) on the first syllable. They were all represented orthographically as two Simplified Chinese characters. The lexical frequency of each real word was controlled within the log frequency range of 3 to 5 according to the SUBTLEX-CH corpus (Cai and Brysbaert 2010). We also evaluated participants' individual level of exposure to and use of Mandarin by querying them through a post-task questionnaire about how frequently they used Mandarin during their childhood and currently in interactions with family, friends, and colleagues.

Results. The Center of Gravity (CoG) was extracted at the mid-point of each fricative. We first compared speakers' [s] productions to their own [ʂ] productions to assess which individuals produced a reliable contrast between the target fricatives. For each speaker, we used a two sample t-test comparing the CoG values of their [s] productions to those of their [ʂ] productions: 17 speakers produced a significantly distinctive contrast and 43 speakers produced an indistinctive (merged) contrast. The productions of these two groups of speakers are visualized in fig. 1. For both merged and distinctive speakers, we employed mixed-effects models to investigate the effects of Fricative and Vowel (both included using deviation coding), as well as their interaction, as fixed factors on CoG values, while accounting for individual variability with by-participant random intercepts. We compared this full model to simpler models excluding one of the fixed effects or their interaction using likelihood ratio tests. For "merged" speakers, only the factor Vowel significantly affected model fit ($\beta = -600.6$, $SE = 43.7$, $\chi^2(1) = 159.6$, $p < 0.001$), while this was not the case for Fricative ($\chi^2(1) < 1$) or the interaction ($\chi^2(1) < 1$). This indicates that the merged group indeed produced fricatives with similar CoG values and that these productions were generally affected by coarticulation with the following vowel. For "distinctive" speakers, the full model was a significantly better fit to the data than models which excluded the factors Fricative ($\beta = -1263.2$, $SE = 98.9$, $\chi^2(1) = 118.5$, $p < 0.001$), Vowel ($\beta = -424.9$, $SE = 98.9$, $\chi^2(1) = 159.6$, $p < 0.001$), and their interaction ($\beta = 632.6$, $SE = 197.9$, $\chi^2(1) = 10.1$, $p < 0.01$). This indicates that in addition to the same coarticulatory effect reported above, for the distinctive group, the CoG values of both target categories were closer in the context of [u] than in the context of [a].

We turned next to social factors that might influence the variation we observed. We examined whether exposure to and use of Mandarin, age group, and gender influenced productions of the target fricatives. Because the CoG values were found to significantly differ according to vowel context, we looked at data from each vowel context separately. We created for each vowel context a linear regression model predicting participants' average CoG differences in the relevant vowel context with individual Mandarin exposure scores, gender, age group, speaker classification (distinctive vs. merged), as well as the interaction between classification and Mandarin exposure score. Our analysis revealed that, for both vowel contexts, speakers with a higher Mandarin exposure score tended to produce a larger contrast between the target fricatives ([a] context: $\beta = 272.9$, $p < 0.01$; [u] context: $\beta = 255.6$, $p < 0.001$). Both models indicated a significant negative effect of being classified as a merged speaker ([a] context: $\beta = -1490.0$, $p < 0.001$; [u] context: $\beta = -816.6$, $p < 0.001$), again reflecting that distinctive speakers produced a larger contrast between the target fricatives. Additionally, the effect of the interaction between a speaker's classification and their Mandarin exposure score on their CoG difference was significant ([a] context:

$\beta = -333.5$, $p < 0.001$; [u] context: $\beta = -188.2$, $p = 0.01$), indicating that for speakers who maintain a distinct fricative contrast, a higher exposure score more strongly correlates with a greater acoustic difference between the target fricatives than for speakers who merge the contrast. This is clearly visible in fig. 2, where the regression lines for merged speakers are nearly flat, in contrast with those for distinctive speakers which show a positive relation between exposure to Mandarin and CoG differences. Concerning the other social factors, none were statistically significant for the [a] context (all $p > 0.05$). For the context of [u], significant effects were observed for gender and age, with male speakers ($\beta = -426.5$, $p < 0.01$) and older speakers [compared to the youngest group] ($\beta = -384.8$, $p < 0.05$) producing less distinct contrasts.

Discussion. In this study, we tested the production of a Mandarin sibilant fricative contrast by bilingual speakers of QSM in two different vowel contexts. The results indicate that both the following vowel and a speaker's Mandarin exposure level are significant predictors of how this contrast is produced. Both "distinctive" and "merged" speakers showed a coarticulatory effect such that CoG values were lower before the vowel [u]. The distinctive group further showed an interaction effect where the significant difference they produced between target [s] and [ʂ] was smaller in the context of [u]. Importantly, speakers with a higher Mandarin exposure score produced greater CoG differences, and we found that this relation was clearly stronger for distinctive speakers than for merged speakers. More work is necessary to explore what other factors influence whether speakers are likely to merge the target contrast or to produce distinctive fricatives and to understand the role of age and gender. Finally, while exposure to and use of Mandarin appeared to relate to how strong of a contrast a speaker was likely to produce (more exposure to Mandarin was correlated with a stronger contrast), the interaction effect suggests that other factors must be at play. What makes a speaker likely to distinguish or merge the contrast in the first place? Future work might benefit from including a measure of acuity alongside the factors explored here.

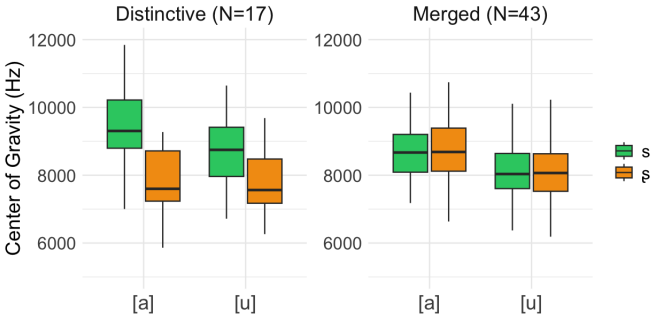


Figure 1: Comparison of CoG value for [s] and [ʂ] fricatives across vowel contexts in L1 QSM Speakers

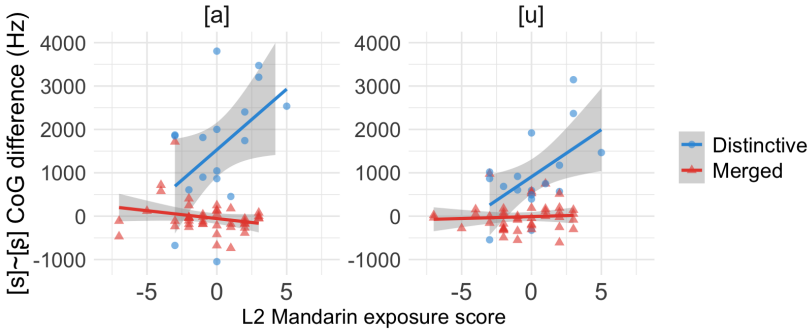


Figure 2: Participants' mean CoG difference as predicted by L2 Mandarin exposure level in each vowel context. More positive scores represent higher exposure to and use of Mandarin compared to QSM; more negative scores represent higher exposure to and use of QSM compared to Mandarin.

References.

Cai, Qing and Marc Brysbaert (2010). "SUBTLEX-CH: Chinese word and character frequencies based on film subtitles". In: *PloS one* 5.6, e10729.

Chang, Yung-Hsiang Shawn and Chilin Shih (2015). "Place contrast enhancement: The case of the alveolar and retroflex sibilant production in two dialects of Mandarin". In: *Journal of Phonetics* 50, pp. 52–66.

Chuang, Yu-Ying and Janice Fon (2010). "The effect of prosodic prominence on the realizations of voiceless dental and retroflex sibilants in Taiwan Mandarin spontaneous speech". In: *Speech Prosody 2010-Fifth International Conference*.

Kubler, Cornelius C (1985). "The influence of Southern Min on the Mandarin of Taiwan". In: *Anthropological Linguistics* 27.2, pp. 156–176.

Lee-Kim, Sang-Im and Yun-Chieh Chou (2022). "Unmerging the sibilant merger among speakers of Taiwan Mandarin". In: *Laboratory Phonology* 13.1, pp. 1–36.

Encoding of speech modes with varying articulatory and phonatory properties; an ERP Investigation

Bryan Sanders[1], Marina Laganaro[1]

[1]1. Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

bryan.sanders@unige.ch, marina.laganaro@unige.ch

Introduction. Speech Motor Control (SMC) refers to the cognitivo-motor ability of encoding, producing and adjusting speech utterances. This ability involves several brain networks (Guenther, 2006, 2016; Scott, 2022) forming a system that exerts voluntary control over the vocal behavior, called "vocal brain" by Belyk & Brown (2017). Voluntary control over speech production has been necessary to adapt to the evolution of vocal communication in humans (Belyk & Brown, 2017). Indeed, control over utterances production allows speakers to modulate their speech to overcome both speaker related (e.g., hearing issues, strong foreign accent in a second language) and contextual related obstacles (e.g., background noise, social rules) in communicative contexts (Tuomainen et al., 2022). In particular, speakers adopt specific mode of speech production referred to as "speech modes" such as loud speech, whispered speech, clear speech and so forth (Zhang & Hansen, 2007). Despite their omnipresence in verbal exchanges, the encoding and the characterization of the brain mechanisms at the origin of different speech modes is still unknown and it is not modelled in current speech production models. Understanding the encoding of speech modes could provide additional insights for instance on the dynamics of activation of the brain regions during the encoding of speech in the Direction Into the Velocities of Articulators (DIVA) model (Guenther, 2016), given that the interaction and the role played by each region during the encoding of speech is still lacking (Tourville & Guenther, 2011). In a previous study, we showed that both loud speech and whispered speech yielded an additional encoding time and different electrophysiological correlates relative to standard speech in a time window preceding the onset of articulation. The EEG/ERP results were however slightly different for loud than for whispered suggesting speech mode specific encoding mechanism (Sanders et al., in prep). In the present experiment, we investigate the behavioral and electrophysiological signatures of two distinct speech modes, namely loud speech and imitation of a foreign accent in contrast to standard speech. Asking participant to fake an accent has never been conceptualized as a speech mode, however it fits the definition of mode of speaking which requires phonatory and articulatory adjustments. We expect that both loud and foreign accent imitation should result in encoding cost in terms of production latency and EEG/ERP differences relative to standard speech. We assume that the electrophysiological signature should be observed in a time period preceding the onset of articulation but possibly not in the same time-window nor on the same brain networks for loud speech and faking an accent, as different adaptations are involved.

Methods. 24 French speakers [22,7 years ± 3,6 years] with no neurological or language impairment were recruited to perform the experiment under EEG recording. They agreed to the form consent and were paid for their participation. In a sound-proof room, participants produced three times a set of 84 bi-syllabic CCVCV French pseudowords (e.g., /priko/) in a delayed production task (a total of 252 productions). The experimental session consisted of three blocks with one speech mode per block: standard speech, loud speech or imitation of an English accent. The latter has been chosen based on results obtained during a pilot assessing the experimental testing of overt speech imitation. The accuracy (ACC) and the latency to initialize speech utterances (RTs) were extracted offline using the CheckVocal Software (Protopapas, 2007). Incorrect (e.g, /gRibu/ instead of /gRibõ/), hesitations (e.g., /bRo/…/bRoga/) or incomplete (e.g., /grad-/) utterances were considered as inaccurate. Two judges did the exact same procedure for ACC and RT resulting in an inter-judge agreement of 95% and 80% respectively. Differences across conditions (speech modes) have been statistically tested with the mixed model approach. In parallel with the production of utterances, high-density EEG was recorded with the 128 Active two Biosemi system (Biosemi V.O.F. Amsterdam, Netherlands). As preprocessing steps, we removed the baseline value (i.e., DC removal), frequencies below 0.1 and above 30 Hz (i.e., high pass and low pass filters), electrical line interference of 50 Hz (i.e., Notch filter). After that, we extracted event-related (ERP) epochs of 350 ms (i.e., 179 TF) that were aligned in a backward manner to the vocal onset of each trial. Furthermore, we matched the number of epochs per conditions, interpolated a maximum of 20 electrodes, changed the reference to the average and applied a spatial filter. Waveform amplitudes, microstates (Michel & Koenig, 2018) and time frequency analyses will be carried out by comparing the standard ERPs to the speech modes ERPs.

Results. Behavioral analyses on the whole sample revealed that accuracy was high overall (M = 96%; Standard Deviation (SD) = 4,65 ; Minimum (Min) = 83% ; Maximum (Max) = 100%) in each condition with no significant difference across speech modes. On average, pseudowords were produced in 507,67 ms (SD = 80,96; Min = 337,72 ms; Max = 707,70 ms) with slight differences across modes that were not statistically significant (p= .054 for standard versus loud speech and p=.132 for standard versus imitation of an accent). The preliminary ERP results on 9 participants are summarized in

Figure 1. Waveform amplitudes analyses (Fig. 1.B) revealed different amplitudes in the last 150 ms preceding vocal onset for each speech mode relative to standard speech, but more extended in terms of electrodes for the accent condition. Results of the Topographic Analysis of Variance (TANOVA) showed that both speech modes had significant time windows of topographic differences in comparison to standard speech (standard-imitation: from -180 to vocal onset; standard-loud: from -100 ms to -80 ms before the vocal onset and from -130 ms to -120 ms). The spatio-temporal segmentation analysis (Fig. 1.A) yielded three stable periods of electrophysiological activity (GEV=93,35) on the averaged ERP signals per condition, with different distribution across the speech modes. We fitted these three template maps into the individual ERPs and run non-parametric statistics on one temporal parameter (the Center of gravity, COG) and one global measurement of occurrence (the Area under Curve, AUC). Differences across modes were observed on COG ($X^2(2)=6.89$, p=.032) but not for the AUC ($X^2(2)= 2.67$, p=.263) on Map A. Map B did not differ across conditions and Map C differed across conditions on the global measurement parameter (COG: $X^2(2)= 9.172$, p=.01 ; AUC: $X^2(2)=6.276$, p=.043).

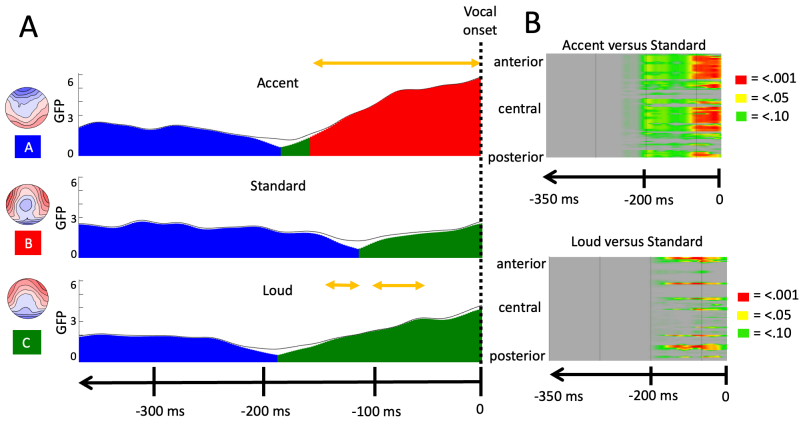


Figure 1: *A. Response-locked spatio-temporal segmentation of the ERPs. Yellow arrows correspond to significant time periods obtained with the TANOVA analysis. B. Pairwise Amplitudes analyses on all electrodes (one per line).*

Discussion. The inialization times across conditions were not statistically different meaning that no encoding cost was necessary to produce speech modes in a delayed speech production task. This result is surprising considering the results observed in the literature (e.g., increased latency of production for speech modes in Zhang & Hansen, 2007) or even previous results on loud speech in our lab (Bourqui et al., in prep). ERP analyses demonstrated that both the faking accent condition and the loud condition yielded an EEG/ERP signature that differed from standard speech in a similar encoding time window preceding to the vocal onset. In these preliminary analyses, the time period showing amplitudes and topographical differences between faking an accent and standard speech seem more robust across the analyses than for loud speech. In other terms, the present results suggest that encoding the parameters for faking an accent and speaking louder than usual will probably be achieved in the same time period likely corresponding to motor programming/parametrization. The EEG/ERP analyses on the whole group should further clarify the differences in time-periods and brain networks between the two modes.

References

Belyk, M., & Brown, S. (2017). The origins of the vocal brain in humans. *Neuroscience & Biobehavioral Reviews*, *77*, 177-193.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of communication disorders*, *39*(5), 350-365.

Guenther, F. H. (2016). *Neural control of speech*. Mit Press.

Koenig, T., & Melie-Garcia, L. (2010). A method to determine the presence of averaged event-related fields using randomization tests. *Brain topography*, *23*, 233-242.

Michel, C. M., & Koenig, T. (2018). EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *Neuroimage*, *180*, 577-593.

Scott, S. K. (2022). The neural control of volitional vocal production—from speech to identity, from social meaning to song. *Philosophical Transactions of the Royal Society B*, *377*(1841), 20200395.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and cognitive processes*, *26*(7), 952-981.

Tuomainen, O., Taschenberger, L., Rosen, S., & Hazan, V. (2022). Speech modifications in interactive speech: effects of age, sex and noise type. *Philosophical Transactions of the Royal Society B*, *377*(1841), 20200398.

Zhang, C., & Hansen, J. H. (2007). Analysis and classification of speech mode: whispered through shouted. In *Eighth Annual Conference of the International Speech Communication Association.*

Velum lowering and tongue tip constriction in German VN sequences across different speech styles

Esther Kunay[1], Lilian von Bressensdorf[1], Philip Hoole[1], Jonathan Harrington[1], Dirk Voit[2], Jens Frahm[2]

[1] *Institute of Phonetics and Speech Processing, University of Munich*

[2] *Max Planck Institute for Multidisciplinary Sciences, Göttingen*

es.kunay|l.bressensdorf|hoole|jmh@phonetik.uni-muenchen.de; dvoit|jfrahm@mpinat.mpg.de

Introduction. Understanding the phonologisation of contrastive vowel nasalisation requires an explanation of how a nasal consonant's two primary articulators - the velum and the nasal consonant's oral constriction - become disassociated from each other over time. For example, during the sound change from Latin *manus* to French *main* [mɛ̃] ('hand'), the vowel became nasalized while the nasal stop was lost. Studies focusing on contrastive vowel nasalisation typically investigate the specific phonetic environments and prosodic factors facilitating velum lowering both in space and time in (C)VN(C) sequences (e.g. Huffman and Krakow 1993; Amelot and Rossato 2007; Solé 1992). Less research has been conducted with respect to the interaction of the velum and tongue tip movements within the same target sequences (Byrd et al. 2009; Oliveira, Martins, and Teixeira 2009). In our study we present data of velum and tongue tip movements obtained from real-time magnetic resonance imaging (MRI) recordings conducted with native speakers of Standard German. As this language exhibits no nasal vowels or strong coarticulatory vowel nasalisation, the articulatory mechanisms fundamental for the production of extensively nasalised vowels can be tested. We are especially interested in whether the tongue tip and the velum are affected differently by varying speaking rates and we intentionally focus on function words, i.e. short and high frequently recurring words in which gestural reduction is likely.

Methods. Real-time MRI data have been acquired for 41 native speakers of Standard German. Here we present results for the first 20 speakers. The speech material used in this study consists of a subset of a larger speech corpus; here we focus on 25 function words ending in either /n/, /ns/ or /nst/ (e.g. *ein, dann, wenns* ('a', 'then', 'if (it)')). The target words were placed in a prosodically unaccented position in varying natural sentences. Words ending in /n/ were followed by a homorganic consonant to avoid assimilation effects. Participants read out the sentences in a moderate (M) and in a very fast (F) speaking mode. Additionally, the target items were embedded in a constant carrier phrase and participants were asked to read the phrases carefully (laboratory style, L). Magnetic resonance images were obtained at 80 fps with in-plane pixel size of 1.41x1.41 mm^2. In addition, synchronic acoustic data were recorded and analysed with respect to segmental boundaries. The images were processed in MATLAB (The Mathworks Inc., 2017, details in Carignan et al. 2021; Kunay et al. 2022), such that time-varying signals were obtained from both the velum and tongue tip movements, from which additional kinematic landmarks were derived to determine the gesture onsets and offsets. Since gestures typically become more reduced and acoustic boundaries become more difficult to determine with increasing speech rate, we decided to extract the signal values of interest as follows: First, the time point of the acoustic boundary in VN was determined for the L condition, i.e. when the vowel and nasal were clearly present in the acoustics. Next, the onsets of the velum lowering and raising gestures across the target word were determined in the L condition. Then, the time point of the acoustic boundary was expressed relative to these onsets of velum lowering and raising. Finally, the signal values of the tongue tip and velum position were extracted at the corresponding relative time points in all conditions.

Results. Data are presented with respect to the extent of spatial gestural reduction at the relative time points described above. There are additional ongoing analyses regarding temporal aspects, i.e. potential re-phasing of the two articulators across the three speech modes; results will be available soon. Data suggest that with respect to the nasal stop, there is much more spatial variation in velum lowering than in tongue tip constriction (TTC) across the different speech rates. **Figure** 1 shows the degree of TTC and velum lowering for the three speaking conditions at the relative time point outlined above (higher values refer to a higher tongue tip and a lower velum position). **Figure** 1 suggests that while for the tongue

tip there is some difference between the L and M conditions on the one hand, and the F condition on the other, by contrast for the velum there are robust differences between all three conditions. This was confirmed by statistical analysis.

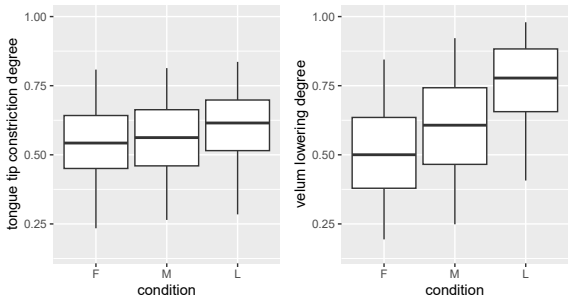


Figure 1: *Tongue tip constriction (TTC) and degree of velum lowering at acoustic VN boundary for L; timepoints for M, F are determined relative to L (see text). Data normalized for each speaker to 0..1 based on 0.5 and 99.5 percentiles across all data in the corpus*

Discussion. The results give insight into the effect of speech rate on the spatial reduction of the tongue tip and velum movement in German nasal stops. Unlike a scenario in which the vowel becomes more nasalised while the oral constriction is reduced (especially in contexts that facilitate reduction), our data suggest that the oral closure was quite stable and velum lowering was decreased in fast speech rate, i.e. closer to velum closure. A higher velum position is not particularly compatible with an increase in nasalisation of the preceding vowel in reduced speech. However, there are actually various outcomes possible when VN sequences change over time. One of them is the emergence of a nasalised vowel and loss of the coda, as in Romance languages (Sampson 1999). Another one is just the loss of the nasal stop without the vowel being necessarily nasalised (Busa and Ohala 1995). The spatial data are more compatible with this strategy, namely a reduction of the nasal gesture in favour of the following oral consonant when the speech condition requires gestural adaptation. Further analysis of gestural phasing will consider whether velum lowering nonetheless aligns earlier with respect to the preceding vowel at the faster speech rates. Moreover, context effects will be taken into consideration, i.e. how the two articulators are affected in longer vs. shorter vowels, in low vs. non-low vowels, and when nasals precede voiceless obstruents vs. voiced sonorants. This study in any case adds to the relatively sparse data comparing reduction patterns in unlinked articulators in hypoarticulated speech that is typical for function words and thus contributes to a better understanding of gestural cohesion. Moreover, it also demonstrates that real-time MRI now not only has the spatial and temporal resolution to address such issues, but also that the analysis methods can readily be applied to large numbers of speakers.

References.

Amelot, Angélique and Solange Rossato (2007). "Velar movements for two French speakers". In: *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken, Germany, pp. 6–10.

Busa, M Grazia and JJ Ohala (1995). "Nasal loss before voiceless fricatives: a perceptually-based sound change". In: *Rivista di Linguistica* 7, pp. 125–144.

Byrd, Dani, Stephen Tobin, Erik Bresch, and Shrikanth Narayanan (2009). "Timing effects of syllable structure and stress on nasals: a real-time MRI examination". In: *Journal of Phonetics* 37.1, pp. 97–110.

Carignan, Christopher, Stefano Coretta, Jens Frahm, Jonathan Harrington, Phil Hoole, Arun Joseph, Esther Kunay, and Dirk Voit (2021). "Planting the seed for sound change: Evidence from real-time MRI of velum kinematics in German". In: *Language* 97.2, pp. 333–364.

Huffman, M.K. and R.A. Krakow (1993). *Nasals, Nasalization, and the Velum*. Phonetics and Phonology. Elsevier Science.

Kunay, Esther, Philip Hoole, Michele Gubian, Jonathan Harrington, Arun Joseph, Dirk Voit, and Jens Frahm (2022). "Vowel height and velum position in German: Insights from a real-time magnetic resonance imaging study". In: *Journal of the Acoustical Society of America* 152.6, pp. 3483–3501.

Oliveira, Catarina, Paula Martins, and António Teixeira (2009). "Speech rate effects on European Portuguese nasal vowels". In: *Tenth Annual Conference of the International Speech Communication Association*, pp. 480–483.

Sampson, Rodney (1999). *Nasal vowel evolution in Romance*. Oxford University Press on Demand.

Solé, Maria-Josep (1992). "Phonetic and phonological processes: The case of nasalization". In: *Language and Speech* 35.1-2, pp. 29–43.

Liquids in Upper Sorbian: an MRI examination

Juliusz Cęcelewski[1], Phil J. Howson[2], Peter Birkholz[3], & Cédric Gendrot[1]

[1]Laboratoire de Phonétique et Phonologie
[2]Leibniz-Zentrum Allgemeine Sprachwissenschaft
[3]Technische Universität Dresden
howson@leibniz-zas.de

Introduction. Liquids are a superclass of segments comprised of two smaller classes: laterals and rhotics. They have been problematic because unlike other classes of segments, it has been difficult to pin down phonetic correlates of the class. Nevertheless, theories connecting the class through articulatory measures have been proposed. For example, Proctor (2011) suggested that the connection between the liquids is that both laterals and rhotics share a coordinated tongue tip/blade and tongue dorsum gesture. The complexity of these segments does typically involve a pharyngeal constriction (Delattre & Freeman, 1968; Alwan, Narayanan, & Haker, 1997; Narayanan, Alwan, & Haker, 1997; Narayanan, Byrd, & Kaun, 1999) accompanied by a more anterior constriction. However, other researchers (e.g., Recasens, 2016) have found that the gestures involved in liquid articulation do not always involve coordination of the tongue tip/blade and dorsum articulators. It is additionally unclear how uvular rhotics would possibly fit into this type of analysis due to the absence of a tongue tip/blade gesture. The purpose of the presented research is to examine the liquids in an endangered language, Upper Sorbian. Upper Sorbian has an alveolar lateral, a uvular rhotic, and a palatalized uvular rhotic. The disparate places of articulation make Upper Sorbian an excellent testing ground to examine the similarities and differences between the liquids. Additionally, few studies have examined 3D volumetric data for uvular rhotics. Based on previous studies, we predict significantly different vocal tract area functions for laterals and rhotics. However, we anticipate similarities in the pharyngeal cavity area functions.

Methods. Four L1 speakers of Upper Sorbian participated in this study (2 male and 2 female). Participants were college-educated, between the ages of 20-24, and had no self-reported history of speech or hearing disorders. Participants were given a list of words with the target segments (/l, ʀ, ʀʲ/) prior to data collection so they could practice producing each segment for an extended duration. Target phonemes were in the word initial position followed by a low central vowel, /a/. MRI data were recorded at the Neuroimaging Center at the Technische Universität Dresden. Data were recorded with a Siemans 3T Trio, with a pixel size of 1.2 mm x 1.2 mm x 1.8 mm. 44 sagittal slices were taken to construct the 3D image of the vocal tract. Participants sustained articulation of a single segment for 14 seconds in order to image their vocal tract shapes. After each trial, MRI images were examined and in cases of blurriness or poor image quality, participants repeated the trial until a clear image was obtained.

3D vocal tract shapes were segmented using ITK-Snap. The automatic segmentation function was used and then visually inspected and corrected by hand. We then extracted 3D vocal tract shapes and opened up the model at the vocal cords (the most posterior section of the model) and at the lip aperture (the most posterior section of the model) using Blender. We did this so that models could be imported into the software VTTF (Echternach, 2016) for area function measurements of the vocal tract. 3D images were also obtained for presentation in Blender. GAMMs were used to compare the area function across speakers using the *mgcv* package (Wood, 2011). We included a smooth term for slice (x-axis) and for the interaction between slice and segment. We included a random by participant intercept with a by-segment slope. Model estimates were extracted with *itsadug* (van Rij et al., 2022) and results were plotted with *ggplot2* (Wickham, 2016).

Results. The GAMM analysis revealed a significant effect for the interaction between slice and /l/ [$F(8.45, 10.56) = 31.77$, $p < 0.001$] and the interaction between slice and /ʀʲ/ [$F(4.40, 5.75) = 2.73$, $p = 0.014$], but not the interaction between slice and /ʀ/ [$F(1, 1) = 1.90$, $p = 0.169$]. The R^2 for the model was 0.6. Figure 1 presents the average area functions for each segment, Figure 2 presents the GAMM differences contours, and Figure 3 presents the example 3D vocal tract shapes for one Upper Sorbian speaker, US04. The results indicated that the lateral in Upper Sorbian differs from both rhotics along several dimensions. The lateral featured a less constricted pharyngeal, uvular, velar, and palatal region, and a more constricted vocal tract geometry in the anterior of the hard palate. The tongue additionally formed a complete constriction along the alveolar ridge. Speakers additionally had lateral side channels on both sides of the tongue. The differences between /ʀ/ and /ʀʲ/ were marginal, but significant. The most posterior section of the pharyngeal cavity was slightly less constricted for /ʀʲ/ than /ʀ/, but the overall shapes of the pharyngeal cavity were otherwise not significantly different. We also observed a more constricted uvular, velar, and palatal region for /ʀʲ/ when compared to /ʀ/. /ʀʲ/ also had a slightly raised tongue blade. The results suggest that the impact of palatalization on the uvular rhotic are marginal but involves slight tongue blade/body raising.

Discussion. We observed significant differences in the vocal tract shapes across the liquids. The lateral shared a small similarity in the lower pharyngeal tract area functions but differed significantly in the middle and upper pharyngeal constriction. There were also no similarities between the laterals and rhotics in terms of the area function along the soft and hard palate. The lack of any striking similarities suggests completely different articulatory configurations for the laterals and the rhotics. The results here leave open the possibility of a link between the two classes in the articulatory

domain with respect to the presence of some type of pharyngeal constriction. However, this research also demonstrates significant disparity in vocal tract area functions and suggests that in fact the basis of the class of liquids may be in the acoustic (Narayanan, Byrd, & Kaun, 1999) or acoustic-perceptual (Howson & Madathodiyil, 2023) domain. One limitation of the study is that the teeth are not visible in the MRI imaging sequence.

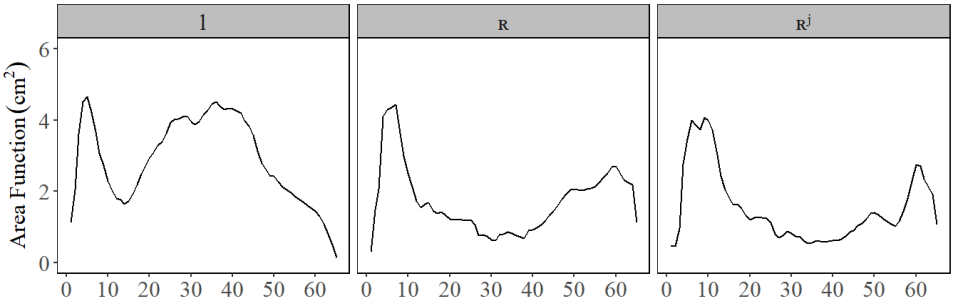


Figure 1: *Average area functions for each segment /l/ (left), /ɹ/ (center), and /ɹʲ/ (right). Measurements are in cm². The x-axis indicates the slice along the vocal tract the measurements were taken from.*

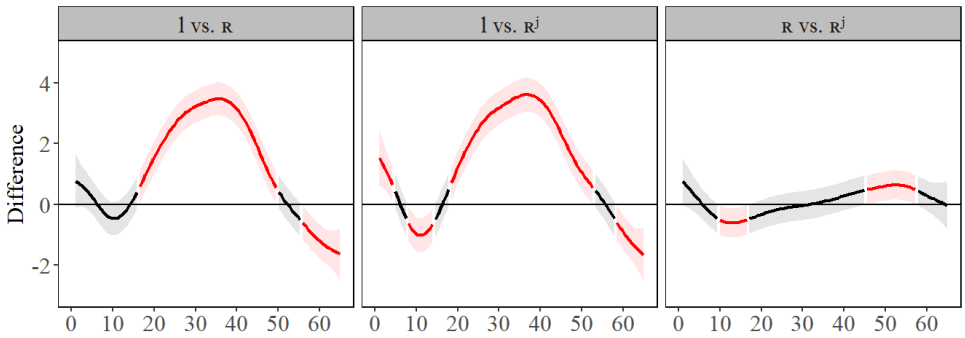


Figure 2: *GAMM difference plots comparing the area functions for /l vs. ɹ/ (left), /l vs. ɹʲ/ (center), and /ɹ vs. ɹʲ/ (right). Red indicates a significant difference between the two area functions.*

Figure 3: *Example 3D vocal tract images for /l/ (left), /ɹ/ (center), and /ɹʲ/ (right) for Speaker US04.*

References

Alwan, A., Narayanan, S., & Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. *Journal of the Acoustical Society of America*, 1079-1089.

Delattre, P. C. & Freeman, D. C. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics*, 29-68.

Echternach, M., Birkholz, P., Traser, L., Flügge, T., Kamberger, R., Burk, F., Burdumy, M., & Richter, B. (2015). Articulation and vocal tract acoustics at soprano subject`s high fundamental frequencies. *Journal of the Acoustical Society of America*, 2586-2595.

Howson, P. J. & Madathodiyil, I. (2023). The cross-linguistic perception of liquids: motivation for the superclass. Speech Communication, 1-8.

Narayanan, S., Alwan, A., & Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPGdata. Part I. The laterals. *Journal of the Acoustical Society of America*, 1064-1077.

Narayanan, S., Byrd, D., & Kaun, A. (1999). Geometry, kinematics, and acoustics of Tamil liquid consonants. *Journal of the Acoustical Society of America*, 1993-2007.

Proctor, M. (2011). Towards a gestural characterization of liquids: Evidence from Spanish and Russian. *Laboratory Phonology*, 451-485.

Recasens, D. (2016). What is and what is not an articulatory gesture in speech production: The case of lateral, rhotic and (alveolo)palatal consonants. Gradus, 24-42.

van Rij, J., Wieling, M., Baayen, R., van Rijn, H. (2022). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. R package version 2.4.1.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society*, 3-36.

Temporal processing of sentence production in Parkinson's disease

Fatemeh Mollaei[1], Alexandra Pool[1], Huw Evans[1]

[1]School of Psychology and Clinical Language Sciences, University of Reading

f.mollaei@reading.ac.uk, alexandra.pool@student.reading.ac.uk, h.evans@student.reading.ac.uk

Introduction. Parkinson's disease (PD) is a multifaceted disorder with motor and non-motor symptoms. Speech deficits are one of the common symptoms with 90% of individuals with PD showing speech production impairments that span prosody, phonation, and articulation subsystem of speech (Duffy, 2019). These deficits can be broken down into two main categories: hypokinetic dysarthria and neurogenic stuttering (Goberman et al., 2010). While there has been more focus on understanding the nature of hypokinetic dysarthria in PD, there has been less focus on the nature of neurogenic stuttering in PD. Understudying the temporal processing of sentence production in PD will inform the nature of neurogenic stuttering in this population. The purpose of this study was to improve our understanding as to which factors determine online, spoken sentence production abilities of adults with PD in sentence production during reading.

Methods. Reading samples of the Rainbow passage were analysed with Praat speech analysis software. Participants comprised thirty-two people with PD as well as thirty-nine neurotypical controls. Durations of pauses that included silent and filled pauses were analysed according to multiple factors. These included (1) the location they occurred: between and within sentences, (2) the syntactic complexity of sentences: simple and complex, and (3) sentence length: number of words. We conducted statistical analysis of general linear models to compare between the two groups.

Results. PD speakers had a significantly greater number of pauses in all variables compared to controls (**Table 1)**. However, only between-sentences and long sentences pauses had greater duration of pause in the PD group. Both sentence complexity and sentence length showed significant effects on the PD and control groups, with longer sentences producing longer and greater number of pauses than shorter sentences. This was seen with complexity, with more complex sentences producing longer and greater number of pauses than simple sentences.

Table 1: *Statistics are reported in means (SDs), duration figures are in seconds*

	PD	Controls
Total number of pauses	38.77 (15.12)	28.05 (1.0)
Between sentence mean duration	0.96 (0.24)	0.77 (0.16)

Discussion. The study provides evidence for the impact of complexity of sentence and sentence length on pause durations in PD speech. This causes a reduction in processing speed during speech production in PD which ultimately affects the planning and programming of sentence production (Salis & DeDe, 2022). This in may be one of the contributing factors to neurogenic stuttering in PD.

References

Duffy, J. R. (2019). Hypokinetic Dysarthria. In J. Duffy (Ed). Motor speech disorders e-book: Substrates, differential diagnosis, and management, Elsevier Health Sciences.165-190.

Goberman, A. M., Blomgren, M., & Metzger, E. (2010). Characteristics of speech disfluency in Parkinson disease. *Journal of Neurolinguistics*, *23*(5), 470-478.

Salis, C., & DeDe, G. (2022). Sentence production in a discourse context in latent aphasia: a real-time Study. *American Journal of Speech-Language Pathology*, *31*(3), 1284-1296.

Subthalamic and cortical encoding of syllable sequences

Andrew M. Meier[1], Alan Bush[2,3], R. Mark Richardson[2,3,4], Frank H. Guenther[1,5,6]
[1]Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA, USA
[2]Brain Modulation Lab, Department of Neurosurgery, Massachusetts General Hospital, Boston, MA, USA
[3]Harvard Medical School, Boston, MA, USA
[4]Department of Brain and Cognitive Science, Massachusetts Institute of Technology,Cambridge, MA, USA
[5]Department of Biomedical Engineering, Boston University, Boston, MA, USA
[6]Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA

Introduction and Methods

In this study, we compared the roles of components in the cortico-basal ganglia-thalamic (CBGT) loop in novel speech sequence production, with a focus on responses in the subthalamic nucleus (STN). Recordings were acquired from electrocorticography (ECoG) and deep brain stimulation (DBS) electrodes in 27 patients undergoing neurosurgery for intractable motor disorders (see Bush *et al*. 2021). ECoG electrodes were located in the left hemisphere, in cortical areas including precentral and postcentral gyrus, inferior and middle frontal gyrus, superior temporal gyrus, and supramarginal gyrus. DBS electrodes were located unilaterally or bilaterally in either the ventral intermediate nucleus of the thalamus or the STN. On each trial of this experiment, subjects listened to triplets of consonant-vowel syllables (e.g. “VI TU GA”) then repeated this sequence aloud. Stimuli were constructed from 12 unique syllables containing combinations of 3 unique vowels and 4 unique consonants. Neural responses were quantified by extracting broadband high-gamma power (70-150hz).

We investigated three types of preferential activity:

1. Preferential activation during production of specific syllable ranks (e.g., high activation only during production of the first or third syllables)
2. Syllable encoding during speech preparation
3. Syllable encoding during production

Response profiles for each electrode were quantified with a 1-way ANOVA comparing mean high-gamma response within the pre-speech or production period for a specific syllable rank. Preparatory encoding was computed by using activity between stimulus offset and before speech onset as the outcome variable and syllable-1 identity as predictor. Syllable encoding was computed for each syllable rank by using response within that speech epoch as outcome variable across trials and syllable identity in that rank as predictor. In order to test for nonrandom spatial distributions of significantly encoding electrodes, we used a chi-square goodness-of-fit test (see Fig. 1A-F insets). Expected frequencies for a given brain area under the null hypothesis, in which significantly encoding electrodes were randomly distributed across areas, were set to the total number of significantly encoding electrodes (α=0.05) multiplied by the proportion of all analyzed electrodes located in the area in question.

Results

Rank selectivity was found at above-chance levels across cortical areas of the speech control network, but not in either STN or thalamus (Fig 1A; chance level indicated by horizontal black line). Surprisingly, preparatory encoding of syllable identity was highest in STN (Fig 1B), seemingly contrary to the frequently described role of the STN in nonspecific global motor inhibition (Jahanshahi *et al*. 2015). Preparatory syllable encoding was also unexpectedly low among cortical regions thought to be involved in speech planning and working memory, i.e. inferior frontal gyrus and middle frontal gyrus) (Liakakis *et al*. 2011), and the majority of cortical electrodes with preparatory syllable encoding were found in sensorimotor cortex (Fig 1C).

During production of the first syllable, STN did not show an above-chance number of syllable-encoding electrodes, while syllable-1 encoding was most prominently found in SMC (Fig 1D). Later in production of the sequence, above-chance levels of syllable encoding were found throughout the CBGT network, including STN (Fig 1E-F). Figure 1G shows the preparatory and early production period of an example DBS electrode, with mean high gamma response timecourses sorted according to the first syllable’s consonant in each trial. This electrode shows preferential ramping activity prior to production onset of syllables containing the consonant /s/, at which point this preferential activity is sharply suppressed.

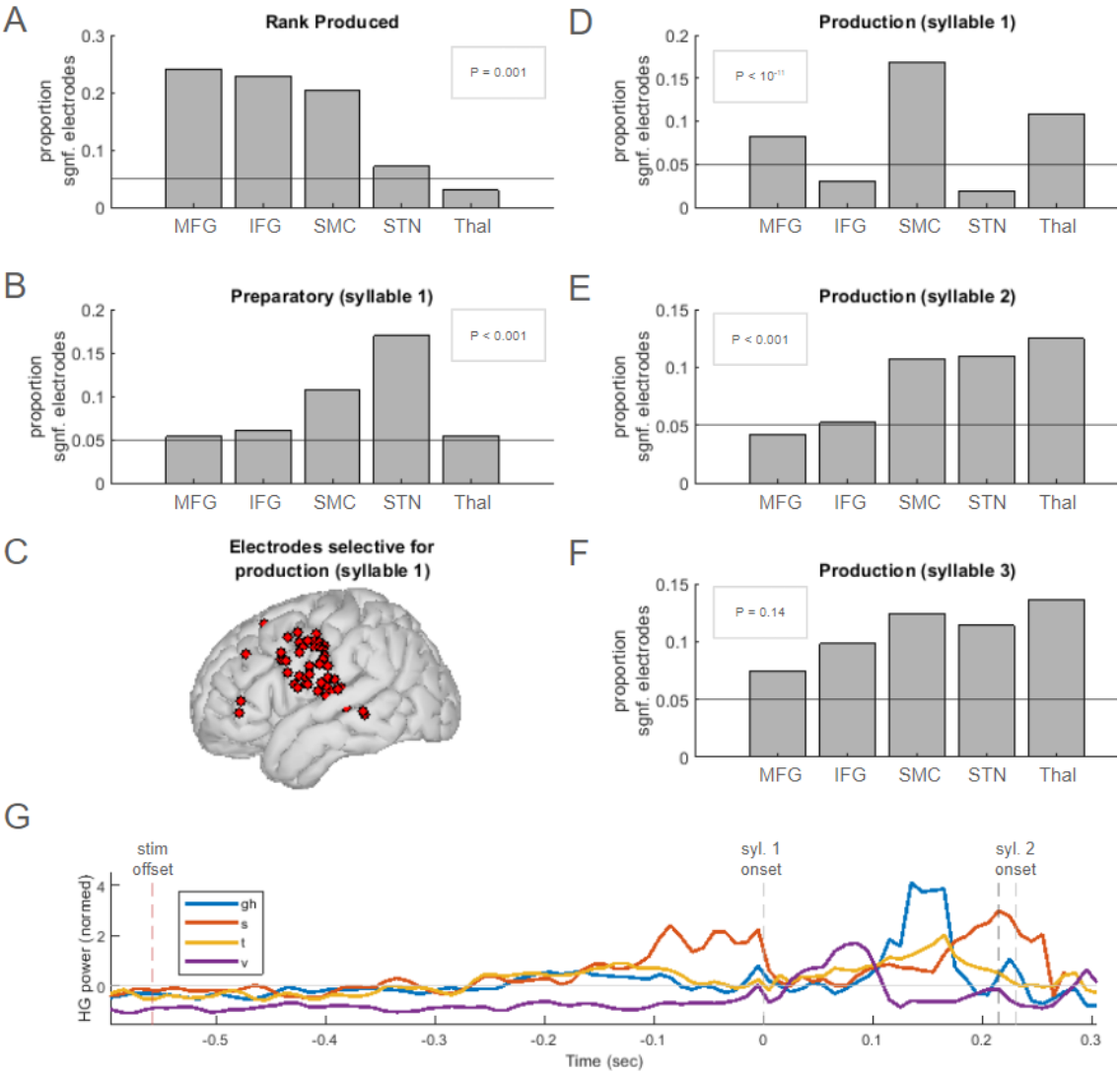


Figure 1. A-B, D-F: Proportions of electrodes significantly encoding rank or syllable identity. (Abbreviations: inferior frontal gyrus (IFG), middle frontal gyrus (MFG), sensorimotor cortex (SMC), subthalamic nucleus (STN), thalamus (Thal).)
C: Cortical locations of electrodes significantly encoding syllable 1 identity during production.
G: High-gamma time-course of example STN electrode.

Discussion

These findings suggest a number of features of feedforward control of speech in the CBGT network, including the STN. First, planning and tracking of ordinal ranks in speech sequences are likely performed through coordination between cortical areas, and not subcortical elements of this network. Second, STN may play an underappreciated role in selection of immediately upcoming speech gestures, through interactions with sensorimotor cortex (more than with higher-order cortical areas). Finally, STN shows higher phonemic selectivity at later parts of the speech sequence, possibly due to early suppression of STN which is necessary for initiating motor output (see Watson & Montgomery 2006).

References

Bush, A., Chrabaszcz, A., Peterson, V., Saravanan, V., Dastolfo-Hromack, C., Lipski, W. J., & Richardson, R. M. (2022). Differentiation of speech-induced artifacts from physiological high gamma activity in intracranial recordings. *Neuroimage*, 250, 118962.

Jahanshahi, M., Obeso, I., Baunez, C., Alegre, M., & Krack, P. (2015). Parkinson's Disease, the Subthalamic Nucleus, Inhibition, and Impulsivity. *Movement Disorders*, 30(2), 128-140.

Liakakis, G., Nickel, J., & Seitz, R. (2011). Diversity of the inferior frontal gyrus - a meta-analysis of neuroimaging studies. *Behavioural Brain Research*, 225(1), 341-347.

Watson, P., & Montgomery Jr, E. B. (2006). The relationship of neuronal activity within the sensori-motor region of the subthalamic nucleus to speech. *Brain and Language*, 97(2), 233-240.

Allophones of Korean /l/: a classification using EMA

Kye Shibata[1], Feng-fan Hsieh[1], Yueh-chin Chang[1]

[1]National Tsing Hua University

kye.shibata@gmail.com, ffhsieh@mx.nthu.edu.tw, ycchang@mx.nthu.edu.tw

Introduction.
This study classifies the allophones of Korean /l/ using data acquired by electromagnetic articulography (EMA). While there have been some attempts at classifying Korean /l/ in previous studies (Crosby and Dalola 2021, Hwang et al. 2019, Lee et al. 2015), no study to our knowledge has utilized EMA.

In this study we will approach the classification of Korean /l/ from multiple angles. First we will look at the lateralization of both sides of the tongue to determine whether the sound is a lateral consonant. Second, we will determine whether there is retroflexion. Finally, we will compare it against Korean /n/, which we will use as a baseline for a non-lateral, non-retroflexed coronal sonorant.

Methods.
Data was collected from three Seoul Korean speakers (more speaker data will be added before the conference) using the Carstens' Articulograph AG501. Our sensor configuration on the tongue follows the "Southern Cross" configuration described in Ying et al. (2021). The speakers were asked to read a mix of Korean words and nonce words in a carrier sentence. We included words that had both /l/ and /n/ in onset and coda positions, in the context of the vowels /a/, /i/, and /u/. The data was processed using MVIEW (Tiede 2005), as well as custom scripts written in MATLAB.

To quantify lateralization, we utilized the lateralization angle method described in Huang et al. (2023), and calculated the average across all trials at a certain point. For onsets, this was at 90% of the entire duration of the onset, and for codas it was at 90% of the duration of the entire syllable.

To examine whether sounds were retroflexed, the angular information provided by the AG501 for each sensor was taken into consideration. Specifically, the elevation angle for the tongue tip (TT) sensor (placed approximately 1cm behind the tongue tip) was examined to determine whether there was a raised tongue tip.

Results.
The lateralization angles calculated from the two parasagittal sensors for /l/ and /n/ in both onset and coda positions are given in Table 1. In the onset position, the speakers had very noticeable differences: Speaker 1 (S1) had similar lateralization for both /l/ and /n/, regardless of the vowel context. Speaker 2 (S2) had greater lateral activity for /n/ when compared to /l/, with a tendency for anti-lateralization in /a/ and /u/ contexts. Speaker 3 (S3) showed right-side lateralization for /l/ only, while /n/ had more symmetric lateral behavior. In coda position, S1 had noticeably greater lateralization on the right side for /bal/ and /bil/, but not /bul/, and had relatively symmetrical, lesser lateralization for the syllables with an /n/ coda. S2 had noticeable left-side lateralization for /bal/, a slight right-side lateralization for /bil/, and relatively symmetric lateralization for /bul/, while syllables ending in /n/ had tended to show significant and symmetric anti-lateralization. S3 once again showed a clear preference for right-side lateralization for all three syllables ending in /l/, and had more symmetry for syllables ending in /n/, though with a slightly more lowered left side.

The angular information provided by the TT sensor, as well as its position within the speaker's mouth, allows us to determine if there was retroflexion during the articulation of the consonant (data from S2 was disregarded as the angular information was not usable due to issues with the sensor placement). Both S1 and S3 had a more posterior place of articulation for /l/ compared to /n/, in both onset and coda positions. In onset position, the TT sensor had a positive elevation angle regardless of vowel context, indicating a raised tongue tip. In coda position, however, the TT sensor only had a positive elevation angle in the context of /a/.

Table 1: *Lateralization angles of /l/ vs /n/ in degrees*

Positive angles indicate a lowering of that side, negative indicates a raising of that side.

	S1		S2		S3	
Syllable	Left	Right	Left	Right	Left	Right
la	1.12	0.36	-1.81	-13.56	7.21	17.85
li	18.27	20.47	4.86	4.55	8.18	24.41
lu	4.04	2.58	-0.41	-1.21	5.06	14.69
bal	7.85	18.43	23.23	-2.53	4.28	25.8
bil	12.84	19.92	1.71	8.68	8.54	15.54
bul	7.77	9.78	5.86	7.78	0.46	10.73
na	2.79	3.32	-9.43	-8.64	11.44	12.13
ni	21.85	22.48	15.99	10.37	15.92	14.06
nu	3.4	3.26	-9.37	-6.92	11.2	16.35
ban	3.99	4.91	-31.66	-28.04	20.59	15.45
bin	8.92	7.29	-29.82	-23.34	26.44	20.48
bun	0.19	-0.16	-31.93	-34.76	26.4	10.72

Discussion.

The results indicate that lateralization for /l/ and /n/ varies based on the speaker: two speakers (S1 and S3) seemed to prefer right-side lateralization, but the degree and context of where this right-side lateralization occurred differed between the two. The third speaker (S2) had a much more varied lateralization pattern, with some contexts having left-preferred lateralization, but not others. S2 also had significant anti-lateralization, something that was not seen in the other two speakers. These differences are likely due to a combination of factors, such as the shape of the speaker's palate.

The coda /l/ in all three speakers had a tendency to have greater asymmetry between the two sides, indicating an asymmetric lateralization such as that observed in Australian English /l/ (Ying et al. 2021). Therefore, the Korean /l/ in coda position is best characterized as a lateral approximant [l]. However, coda /l/ had a more posterior place of articulation when compared to /n/, which appeared to be more dental. Additionally, TT raising was observed in the context of /a/, warranting its classification as a retroflex lateral [l].

In onset position, there was significant variation across speakers. While S1 and S3 indicated TT raising for all onset /l/'s, lateralization was only observed in S3. S2 had inconsistent lateralization in onset position, and had high variance between trials. It is difficult to provide a generalization that covers all speakers, but we may conclude that it is a retroflexed approximant with optional lateralization.

References

Crosby, D., & Dalola, A. (2021). Phonetic variation in the Korean liquid phoneme. *Proceedings of the Linguistic Society of Americ*a, 6(1), 701-712.

Huang, J., Shibata, K., Hsieh, F. F., Chang, Y. C., & Tiede, M. (2023). The L~N merger in Southwestern Mandarin: An articulatory study. Presented at *The 20th International Congress of Phonetic Studies* in Prague, Czech Republic.

Hwang, Y., Charles, S., & Lulich, S. M. (2019). Articulatory characteristics and variation of Korean laterals. *Phonetics and speech sciences*, 11(1), 19-27.

Lee, Y. J., Goldstein, L., & Narayanan, S. S. (2015). Systematic variation in the articulation of the Korean liquid across prosodic positions. In *ICPhS*.

Tiede, M. (2005). MVIEW: software for visualization and analysis of concurrently recorded movement data. New Haven, CT: Haskins Laboratories.

Ying, J., Shaw, J. A., Carignan, C., Proctor, M., Derrick, D., & Best, C. T. (2021). Evidence for active control of tongue lateralization in Australian English/l. *Journal of Phonetics*, 86, 101039.

A framework for modeling the rhythmic organisation of speech and the impact of perceptual cues on production

Mélen Guillaume[1,2], Julien Diard[2], Anahita Basirat[1]

[1] *Univ. Lille, CNRS, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives, F-59000 Lille, France*

[2] *Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France*

melen.guillaume@univ-lille.fr

Introduction. Despite a growing body of evidence about inter-speaker alignment, on the one hand, and the impact of musical stimulation on language processing on the other hand, these phenomena are strikingly underrepresented in computational and neurobiological models of speech perception and production. While not designed to specifically account for these phenomena, computational approaches in speech motor control present relevant characteristics that can address at least some facets of these phenomena, including those related to sequential and temporal aspects. According to the neurocomputational GODIVA model (Bohland, Bullock, and Guenther 2010; Guenther 2016), the sequential and temporal aspects of speech are coordinated by projections from the pre-supplementary motor area to the supplementary motor area and by interactions via subcortical loops, mainly including the basal ganglia. Despite the precision about the brain circuits involved, the computations which underlie the control of rhythmic aspects of speech are less specified in the model compared to those related to the control of segmental aspects.

Another aspect which has received less attention in speech production models concerns the influence of external stimuli perception on speech production. For instance, typical individuals, but also individuals with Parkinson's disease (PD), entrain to the temporal aspects of others' speech (Späth et al. 2022). In this study, the convergence of articulation rate and that of the temporal organization of the perceptual center of the syllable were examined. Results showed that PD participants synchronize more than controls to a model sentence, that is to say, the rate and rhythm of their produced sentences converge more to those of the model sentence. Interestingly, the timing of music listened to before speech production can also affect the pace of speech, with slower musical cues leading to slower speech production (Jungers and Hupp 2018). However, this transfer from perception to production has not been thoroughly explored.

The present study seeks to fill this gap through a computational approach. More precisely, our overall objective is therefore to develop a computational model of speech production able to account for effects of prior perceptual cues on production.

Methods.

A first step towards this objective consists in building a probabilistic fusion model that combines external cues from the processing of prior stimuli in a speech planning sequence. To do so, we will build on two existing frameworks. First, the COSMO family of models (Laurent et al. 2017) was developed to propose integrative architectures linking speech perception and production in a unified probabilistic framework. Second, the latest iteration in this model family, the COSMO-Onset model (Nabé, Schwartz, and Diard 2021; Nabé, Schwartz, and Diard 2022), contained mechanisms for the fusion of information about temporal segmentation of acoustic signal according to syllabic events, with the fusion of both bottom-up cues from the acoustic signal, on the one hand, and top-down cues from lexical and prosodic knowledge, on the other hand.

More precisely, we develop a probabilistic model composed of two components: the adaptive temporal controller and the integrator. The temporal controller focuses on syllables as their basic linguistic units, and plans their temporal organization, according to syllable identity and prosodic constraints, such as the accentual patterns of French. For instance, the adaptive controller is able to control the speech rate and the lengthening of the last syllable of the word, to mark its prosodic prominence. The integrator allow the probabilistic fusion of information between perceptual cues and the speaker's intrinsic timing. The integrator modulates the parameters of the temporal controller, such as the speed at which a syllable is deactivated to make place for the following syllable, as a function of the perceptual input.

The second step consists in choosing a computational framework to implement this temporal fusion mechanism. We chose the GODIVA model framework for two reasons. First, GODIVA allows planning and realization of sequences of

syllables, with precise control of their temporal arrangement. Therefore, it may be adapted so that this arrangement is affected by external information, such as perceived rhythms. Second, GODIVA proposes neuroanatomical correlates of its components, which is particularly relevant to model speech production in neurological pathologies such as PD.

Results. We present preliminary modeling results, with a formal study of fusion models of temporal information in the context of speech planning. More precisely, we define variants of probabilistic fusion models, according to two dimensions. First, we consider temporal planning models that either deal with the probability that there is a syllabic event at time t, or with the probability distribution over time for the next syllabic event. Second, we also consider fusion models with different mathematical properties, such as multiplicative probabilistic combination (resulting in computing temporal compromises of low uncertainty) or additive probabilistic combination (resulting in maintaining temporal alternatives, thus resulting in higher uncertainty). Our results are thus both the mathematical definitions of several variants of the integrator component, and computed simulations to study their resulting properties.

Discussion. The presented results propose a first, preliminary step towards our overall objective. This study of various alternative models of probabilistic fusion of temporal information paves the way towards their integration into the GODIVA model. This will allow the simulation of several model variants, for the formal comparison of their ability to account for experimental data. This will shed light on plausible neurocognitive mechanisms for the integration of perception and intrinsic production cues by speakers during speech sequence planning. This has potential impact on the understanding of speech production pathologies, such as Parkinson's disease or stuttering. For such pathologies, experimental evidence suggest a critical role of interactions between perceptual processing of speech or musical cues and subsequent speech planning performance.

References.

Bohland, Jason, Daniel Bullock, and Frank Guenther (July 2010). "Neural Representations and Mechanisms for the Performance of Simple Speech Sequences". In: *Journal of Cognitive Neuroscience* 22.7, pp. 1504–1529. DOI: 10.1162/jocn.2009.21306.

Guenther, Frank (2016). *Neural control of speech*. Mit Press.

Jungers, Melissa K and Julie M Hupp (2018). "Music to my mouth: evidence of domain general rate priming in adults and children". In: *Cognitive Development* 48, pp. 219–224.

Laurent, Raphaël, Marie-Lou Barnaud, Jean-Luc Schwartz, Pierre Bessière, and Julien Diard (2017). "The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception". In: *Psychological Review* 124.5, pp. 572–602.

Nabé, Mamady, Jean-Luc Schwartz, and Julien Diard (2021). "COSMO-Onset: a neurally-inspired computational model of spoken word recognition, combining top-down prediction and bottom-up detection of syllabic onsets". In: *Frontiers in Systems Neuroscience* 15, p. 653975.

— (2022). "Bayesian gates: a probabilistic modeling tool for temporal segmentation of sensory streams into sequences of perceptual accumulators". In: *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. Ed. by J. Culbertson, A. Perfors, H. Rabagliati, and V. Ramenzoni, pp. 2257–2263.

Späth, Mona, Ingrid Aichert, Dagmar Timmann, Andrés O Ceballos-Baumann, Edith Wagner-Sonntag, and Wolfram Ziegler (2022). "The role of the basal ganglia and cerebellum in adaptation to others' speech rate and rhythm: A study of patients with Parkinson's disease and cerebellar degeneration". In: *Cortex* 157, pp. 81–98.

Impact of Boxing on Vowel Articulation and Self-Assessment of Speech Production by Parkinson's disease Speakers

Caroline Menezes, Beth Ann Hatkevich

University of Toledo

Caroline.menezes@utoledo.edu, bethann.hatkevich@utoledo.edu

Introduction. Parkinson's disease (PD) is a disease of the central nervous system that results in motor and non-motor deficits (Golbe, Mark, & Sage, 2010). Affecting ten million people worldwide, Parkinson's disease occurs due to a loss of dopaminergic neurons in the brain, specifically within the substantia nigra (Radomski & Latham, 2014; Parkinson's Disease Foundation, 2017). Some symptoms have a direct influence on the person's ability to participate in social occupations. Due to motor and non-motor declines in individuals with PD, an individual may experience difficulties in many areas of their life including their ability to speak. Researchers estimate that 89% of people with PD have a speech or voice disorder (Ramig et al., 2008). An individual's ability to enunciate clearly is altered as Parkinson's disease progresses. It is important to note how these changes can impact areas of an individual's life. As speech intelligibility decreases and other communication skills show decline, the quality of life of individuals with PD can be significantly impaired (Streifler & Hofman, 1984; Fujii & Wan, 2014). Individuals with PD appear less interested, less friendly, less involved, and less happy than peers of the same age that do not have Parkinson's disease (McGill University, 2010). Negative perceptions regarding the communication abilities of individuals with PD can limit social occupations and opportunities and may prohibit individuals from wanting to participate in meaningful, social occupations (McGill University, 2010).

A newer area of research suggests that exercise may be able to improve the symptomology of individuals with Parkinson's disease, which could also lead to improvements in the speech mechanism (Shu, Yang, Yu, Huang, Jiang, Gu, & Kuang, 2014). Bradykinesia, one of the cardinal features of Parkinson's disease, can be reduced through exercise (King & Horak, 2009). This reduction in bradykinesia has the potential to have a major influence on the disordered speech of individuals with PD, especially when considering the potential link between akinesia of the limbs and akinesia of the speech mechanism (Rusz, et al., 2016). King and Horak (2009) suggests boxing exercises with individuals with PD, due to the complex nature and incorporation of whole-body movements that boxing encompasses. There is a lack of literature regarding the direct benefits that boxing has on Parkinson's disease, as well as a lack of evidence regarding the direct benefits that exercise may have on the speech mechanism and social engagement of individuals with Parkinson's disease. This study is a preliminary step to investigate speech production and the impact it may have on the social occupations of a person with PD.

Methods. This is a mixed-methodology study designed to assess the effect of therapeutic boxing on the articulation of speech in persons with Parkinson 's disease and compare that to the patient's self-report. Four male individuals with PD from a local boxing club participated in this study. All individuals were considered to be moderately impacted by the disease. All subjects spoke American English. All participants displayed tremors and stiffness.

Stimuli consisted of six American English vowels inserted in the carrier phrase "Please say b[target-vowel]d again" similar to the seminal Hillenbrand study (Hillenbrand et al., 1995). Subject read each utterance six times.

The experiment was a prebox-postboxing paradigm that evaluated both speech and reach skills. The experiment used a randomized block design with some participants recording speech first and reach tasks second and vice versa. The experiment was recorded in two days separated by a week. Between the speech and reach tasks individuals participated in a 30 minutes structured boxing session. When the experiment was completed subjects answered a survey that discussed the benefits of boxing therapy on their speech production.

Speech recordings were made using the Electromagneto Articulograph AG 501. Three reference coils were placed on the left and right mastoid process and on the gingival surface between the maxillary incisors. Coils were also attached to the upper and lower lips, the mandible, the tongue blade, dorsum and tip all aligning on the sagittal plane. For this paper only, the acoustic data is reported for the lack of space. All acoustic data were labelled and analysed using the PRAAT software (Boersma & Weenink, 1992–2022). A PRAAT script was used to extract the first and second formant at the centre of the target vowel. These values were then used to compare vowel spaces of the subject's pre and post-boxing speech. We hypothesize that larger vowel space in the post-boxing condition would indicate a facilitation effect of boxing, but a reduced vowel space would indicate that the subjects were fatigued.

Results. The measure of the effect of therapeutic boxing was studied through formant analysis of four cardinal vowels and the inherent vowel space they create (Robertson & Hammerstadt, 1996; Sapir, 2014). Hyperarticulation would result

in larger vowel spaces hypoarticulation. If boxing has a facilitatory effect, the vowel spaces in the post-boxing condition would be larger than the pre-boxing condition. The reverse would be expected if the subjects were fatigued from the boxing practice.

All speakers showed individual differences. The speakers with less than one year (3-9 months) of diagnosis manifested clear separation of the cardinal vowels and lack of vowel centralization. Furthermore, both speakers showed that the post-boxing vowel space is larger than the pre-boxing condition, indicating that boxing had a facilitatory effect on speech articulation. However, for one subject, there was generally no significant difference (paired T-scores) between pre and post-boxing vowel formants for the cardinal vowels. The speaker with 2.5 years of diagnosis revealed a vowel space that was completely reduced such that his high back vowel was produced closer to his low front vowel. There is still clear separation between high front vowel /i/ and the low back vowel /ɑ/. Following boxing his vowels are more distinct and his vowels space is much expanded. The speaker living with PD for the longest period (3 years) showed a negative effect of boxing on vowel intelligibility. However, Paired Sample T-test showed no significant difference between the pre and post-boxing formant values.

Table 1: *Paired T-scores for tense and lax vowels that were significantly different.*

Subject			Paired Differences					t	df	Sig. (2-tailed)
			Mean	Std. Deviation	Std. Error Mean	95% CI Lower	95% CI Upper			
PD 006	æ	Pre F2 - Post F2	-405.97	151.50	61.85	-564.96	-246.98	-6.56	5	.001
	i	Pre F1 - Post F1	36.90	13.08	5.34	23.18	50.62	6.91	5	.001
		Pre F2 - Post F2	-83.05	42.83	17.49	-128.00	-38.10	-4.75	5	.005
	ɪ	Pre F1 - Post F1	36.07	22.62	9.23	12.33	59.80	3.91	5	.011
	u	Pre F2 - Post F2	225.32	165.08	67.39	52.08	398.56	3.34	5	.020
PD 005	ɑ	Pre F2 - Post F2	72.78	44.12	18.01	26.48	119.08	4.04	5	.010
	i	Pre F2 - Post F2	73.08	24.86	10.15	46.99	99.16	7.20	5	.001
	ʊ	Pre F2 - Post F2	92.11	79.13	32.31	9.07	175.16	2.85	5	.036
PD 009	ɪ	Pre F1 - Post F1	-50.58	35.36	14.44	-87.68	-13.47	-3.50	5	.017

All subjects expressed positive feelings about the community based therapeutic boxing. They claim it increased their ability to socially engage except for one speaker, who was diagnosed with PD only three months prior to this recording. This speaker also claimed that his speech did not benefit from boxing. However, his vowel space shows the largest facilitation following a short boxing routine. But it is clear that his perception about his speech holds him back from being socially comfortable. On the other hand the speaker who was fatigued after boxing, was not sure if boxing helped his speech but was very positive about his social comfort level. This preliminary study appears to tap into the importance of how one feels to be able to speak better and thereby engage more socially.

Discussion. This study had a mixed results due to the influence of severity of the disease. While all four subjects were determined to be moderately impacted by the disease we see that the longer one has been diagnosed with the disease the more their speech is affected. But in the same study we also see that patients with Parkinson's engage in their environment and speak better based on their perception of well being.

References

Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Fujii, & Wan, C. Y. (2014). The role of rhythm in speech and language rehabilitation: The SEP hypothesis. *Frontiers in human neuroscience*, 8(777). doi:10.3389/ fnhum.2014.00777

Golbe, L. I., Mark, M.H., and Sage, J.I. . (2010). Parkinsons Disease Handbook, 1-32. Retrieved from http://www.apdaparkinson.org/userfiles/files/PDHBRev09Repr10.pdf

King, L. A., & Horak, F. B. (2009). Delaying mobility disability in people with Parkinson disease using a sensorimotor agility exercise program. *Phys Ther*, 89(4), 384-393. doi:10.2522/ptj.20080214

McGill University (2010). Parkinson's disease research uncovers social barrier. *ScienceDaily*. https://www.sciencedaily.com/releases/2010/02/100202120815.htm

Parkinson's Disease Foundation, Inc. (2017). Statistics on Parkinson's. Retrieved from http://www.pdf.org/en/parkinson_statistics

Radomski, M. V., & Latham, C. A. T. (2014). *Occupational therapy for physical dysfunction (7th edition)*. Philadelphia: Lippincott Williams & Wilkins.

Ramig, L. O., Fox, C., & Sapir, S. (2008). Speech treatment for Parkinson's disease. *Expert Review of Neurotherapeutics,* 8(2), 297-309. doi:10.1586/14737175.8.2.297

Robertson L,T, & Hammerstadt J.P. (1996) Jaw movement dysfunction related to Parkinson's disease and partially modified by levodopa. J Neurol Neurosurg Psyhiatry 60:41–50

Rusz, J., Tykalová, T., Krupička, R., Zárubová, K., Novotný, M., Jech, R., . . . Růžička, E. (2016). Comparative analysis of speech impairment and upper limb motor dysfunction in Parkinson's disease. *Journal of Neural Transmission*, 1-8. doi:10.1007/s00702-016-1662-y

Shu, H. F., Yang, T., Yu, S. X., Huang, H. D., Jiang, L. L., Gu, J. W., & Kuang, Y. Q. (2014). Aerobic exercise for Parkinson's disease: a systematic review and meta-analysis of randomized controlled trials. *PLoS One,* 9(7), e100503. doi:10.1371/journal.pone.0100503

Streifler, M., & Hofman, S. (1984). Disorders of verbal expression in parkinsonism. *Adv Neurol*, 40, 385-3

Dang, J., & Honda, K. (2002). Estimation of vocal tract shapes from speech sounds with a physiological articulatory model. *Journal of Phonetics*, 30(3), 511-532.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W. Hardcastle & A. Marchal (Eds.), Speech production and speech modelling. Dordrecht: Kluwer Academic. 131-149.

Mermelstein, P. (1973). Articulatory model for the study of speech production. The Journal of the Acoustical Society of America, 53(4), 1070-1082.

Perrier, P., Payan, Y., Zandipour, M., & Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *The Journal of the Acoustical Society of America*, *114*(3), 1582-1599.

Orofacial somatosensory abilities in speech motor control: a conceptual framework

Jean-François Patri[1], *Gilles Vannuscorps*[1]

[1] Psychological Sciences Research Institute, Université catholique de Louvain, Belgium

jean-francois.patri@uclouvain.be, gilles.vannuscorps@uclouvain.be

Introduction. Speech production is a highly complex motor task, requiring precise timing and coordination of orofacial muscles. While it is currently acknowledged that both auditory and somatosensory information play a crucial role in speech motor control (Guenther, 2016), many open questions remain on their specific and relative contributions. The last decades of research have provided many insights into the involvement of several key features of the auditory stream for the control of speech, highlighting e.g. the link between perceptual auditory acuity and the ability to produce speech with greater precision and contrast (Perkell, 2012), the disruptive effect of delays (Stuart et al., 2002), the effect of its deprivation or masking (Lane & Tranel, 1971), or the compensation and adaptation to sustained online perturbation of pitch (Elman, 1981) and formants (Munhall et al., 2009). In parallel, research on the role of oral somatosensation in speech production has developed at a slower pace, progressively unfolding from the debate on its actual involvement in speech motor control (Tremblay et al., 2003), to the evaluation of the effect of its integrity on different aspects of speech proficiency by using a variety of tools and methods. Currently, the empirical and theoretical research on the role of oral somatosensory processing in speech production remains largely underdeveloped as compared to the attention given to audition. More specifically, it remains unclear what the key somatosensory abilities required for speech production are, what indexes and methodologies to use for their assessment, as well as how information from auditory and somatosensory streams is eventually integrated. Consequently, whether, and if so how, oral somatosensory deficits may contribute to different speech disorders, such as stuttering, apraxia of speech and dysarthria remains a largely unresolved question.
One fundamental factor accounting for the disparity in research attention between auditory and somatosensory aspects of speech is clearly empirical: the experimental assessment and manipulation of auditory information is currently much more accessible and less challenging than probing oral somatosensation. However, an additional reason could be the current lack of a comprehensive framework that explicitly identifies and distinguishes between different types of oral somatosensory abilities, as well as their specific functional involvement in speech motor control. We believe that such a conceptual framework would be valuable for the community as it would make it easier to gather, structure and communicate about the existing variety of research directions and findings. It would further support the specification of new research questions, leading to both theoretical and empirical advances for the understanding of the causes and consequences of sensory processing dysfunction in speech disorders, as well as for their assessment and treatment by speech pathologist. The aim of this work is to progress in this direction. We will propose a tentative framework for the organization of research questions, methods, and tools on the role of somatosensation in speech production. We begin by exposing the framework and then use it to organize existing research questions and tools.

Methods. We adopt a theoretically driven approach in which we propose to identify specific somatosensory abilities in reference to the set of somatosensory processes involved in speech motor control. To highlight these sensory processes, we follow the global architecture of current models of speech motor control, distinguishing on the one hand what are the specific goals defining the speech motor task, and on the other hand what is the cascade of control processes required to achieve them. By doing so we hope to make a clearer emphasis on the definition of specific somatosensory abilities with respect to specific speech motor goals. Figure 1 gives a schematic representation of the proposed framework.
The goals of the speech motor task correspond in essence to how current models of speech motor control decompose the flow of speech gestures in terms of sensory or articulatory targets. In the DIVA model, these are represented by the so-called sensory target maps. In the Task-Dynamic model these would correspond to the so-called gestural scores. Without diving into the specificities of these frameworks, we consider a broad characterization of motor goals along three conceptual axes - related to space, time, and dynamics/forces respectively - and this within each organ of the vocal tract, such as to specify respectively the required spatial configuration of the considered event (e.g. a specific tongue posture, or the degree of constrictions), the temporal features specifying the moment, order and duration of these events, and the dynamic features specifying the transitions between events, but also the amount of effort in holding a posture or a constriction, or still the degree of tension modulating the vibratory modes of the vocal folds.
Sensory processes correspond to the hierarchy of processing steps that go from the decoding of raw sensory signals (e.g. neural spikes from mechanoreceptors or muscle spindles) to their translation into higher levels of sensory representations that are exploited by the motor system for the selection of motor commands in task space. However, although the relevant levels of sensory processing steps involved in speech have been relatively well established for audition (e.g. from cochlear signals to formant spaces for vowels), what the specific levels of oral somatosensory representations are for speech production remains largely unknown. Current models of speech motor control directly identify somatosensory space with

articulatory configurations (either articulators' positions, contact or constriction locations), or at best with muscle lengths (Parrell et al., 2019), but in neither case it is truly known whether and if so, how sensory processes actually derive such somatosensory features from the lowest levels of somatosensory signals. Acknowledging the inherent complexity of this question, we propose to begin by identifying somatosensory abilities in a broad sense along the same conceptual axes formulated above for the characterization of motor goals, and further declining them into more specific abilities. Hence, along the spatial axis we specify somatosensory abilities by considering different aspects of spatial resolution (with respect to both tactile and proprioceptive modalities) and of spatial localization for tactile events. Along the temporal axis, we consider aspects of temporal resolution (e.g. distinguishing equal vs unequal durations, as well as synchronous vs asynchronous events), and aspects of delays in sensory integration, both for inter-articulatory and inter-sensory events. Finally, along the dynamic/force axis, we consider the question of sensitivity thresholds for tactile events, and of resolution for the discrimination between levels of effort or contact pressures. Empirically, a variety of approaches have been proposed to assess some of the above-mentioned abilities. These approaches can be classified into an afferent or perception-based perspective and an efferent or production-based perspective. We propose this framework as an effort to organize them and discuss their relevance in targeting specific speech related somatosensory processes.

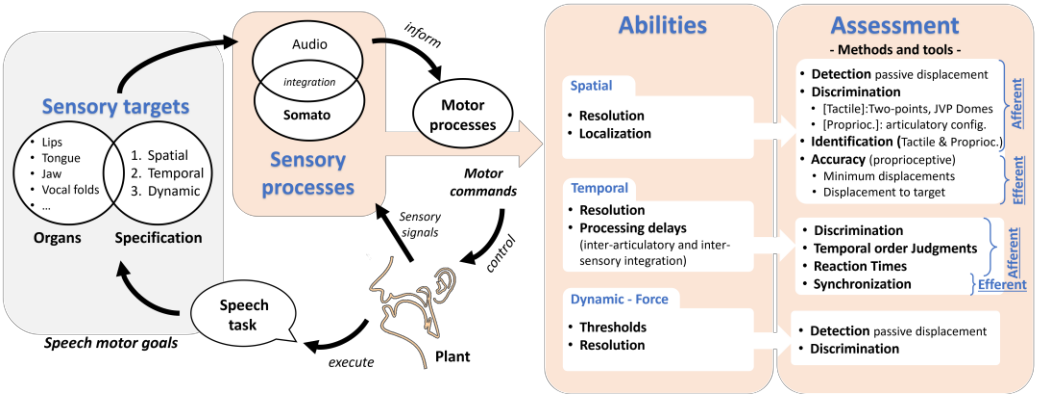


Figure 1*: Schematic representation of the proposed framework for the specification and assessment of somatosensory abilities in speech motor control.*

Results. We apply the current framework to organize and discuss the different approaches that have been conducted to explore the role of oral somatosensory integrity in stuttering. We observe that studies have mainly focused on spatial somatosensory abilities, but that little or no attention has been given to temporal or force-related aspects. In an effort to begin bridging this gap we will provide preliminary data on three tentative methodologies proposed to assess temporal aspects of oral somatosensory processing in stuttering: reaction times for tactile vs auditory stimuli, inter-articulatory temporal order judgements (tactile-tactile events), and inter-sensory temporal order judgements (auditory-tactile events).

Discussion. This paper addresses the imbalance in research focus between auditory and somatosensory information in speech production. We propose a tentative framework to elucidate the specific oro-facial somatosensory abilities crucial for speech motor control by aligning them with specific speech motor goals, considering spatial, temporal, and force-related dimensions and integrating insights from current models of speech motor control. We hope that the proposed framework will be valuable not only for organizing the current landscape of research into oral somatosensation in speech but will also prove useful for future empirical and theoretical advancements, facilitating a deeper understanding of the role of oral somatosensation in speech disorders and supporting the development of effective assessment and treatment strategies by speech pathologists.

References

Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *The Journal of the Acoustical Society of America*, *70*(1), 45–50.

Guenther, F. H. (2016). *Neural control of speech*. The MIT Press.

Lane, H., & Tranel, B. (1971). The Lombard Sign and the Role of Hearing in Speech. *Journal of Speech and Hearing Research*, *14*(4), 677–709.

Munhall, K. G., MacDonald, E. N., Byrne, S. K., & Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *The Journal of the Acoustical Society of America*, *125*(1), 384–390.

Parrell, B., Lammert, A. C., Ciccarelli, G., & Quatieri, T. F. (2019). Current models of speech motor control: A control-theoretic overview of architectures and properties. *The Journal of the Acoustical Society of America*, *145*(3), 1456–1481.

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics*, *25*(5), 382–407.

Stuart, A., Kalinowski, J., Rastatter, M. P., & Lynch, K. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. *The Journal of the Acoustical Society of America*, *111*(5), 2237–2241.

Tremblay, S., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature*, *423*(6942), 866–869.

Mapping Speech and Facial Muscles: Using Generalized Additive Modeling to Understand Speech Production through Electromyography

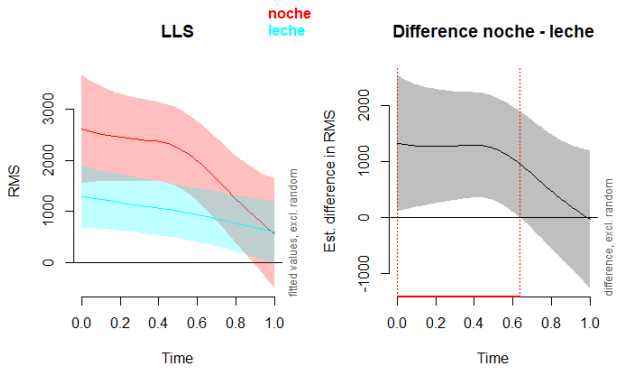
Inge Salomons[1], Inma Hernáez[1], Eva Navas[1], Martijn Wieling[2]

[1]*HiTZ Basque Center for Language Technology, University of the Basque Country (UPV/EHU), Spain*
[2]*University of Groningen (UG), The Netherlands*
inge.salomons@ehu.eus, inma.hernaez@ehu.eus, eva.navas@ehu.eus, m.b.wieling@rug.nl

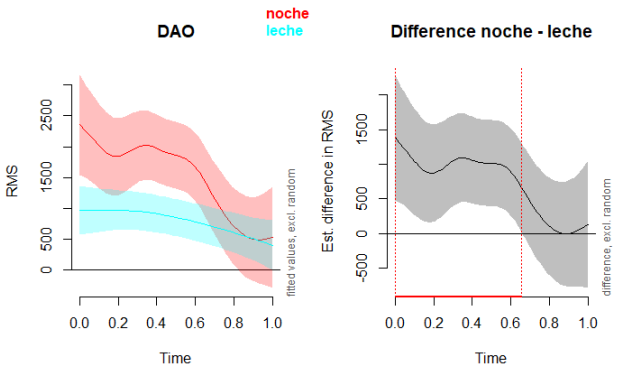
Introduction.
Speech production is a complex human communication tool, requiring the involvement of the brain, lungs, larynx, muscles, and mouth. If any of those components do not function properly, producing speech naturally can become challenging, if not impossible. This is the case for people who have undergone a laryngectomy, a surgery to remove the larynx. They rely on alternative communication methods that do not require the use of the lungs and the vocal cords. An example of such a method is a technological approach, in which information from the facial muscles while speaking silently (i.e. mouthing) is used to predict the speaker's intended words, after which the predicted speech is produced synthetically (Gonzalez-Lopez et al., 2020). To retrieve this information from the muscles, electromyography (EMG) is used. In short, a raw EMG signal represents the activity of motor unit action potentials, and the signal amplitude relates to the degree of muscle activation and force. The activation pattern is most reliably represented with the root mean square (RMS; Vojtech and Stepp (2021)). For the technological approach, first a database of EMG and simultaneous audio signals while speaking audibly (by typical speakers) or silently (by typical and laryngectomized speakers) is created. Subsequently, this database is used to train a machine-learning model with audible data from typical speakers, which is finally evaluated with silent data from laryngectomized speakers. For English, several efforts have been made towards this approach (Diener et al., 2020; Gaddy & Klein, 2020; Wand et al., 2014). Hernáez et al. (2022) introduced a project to develop this technology for Spanish speakers, of which we use the database for this study. So far, the highest accuracy we achieved when predicting the phone (out of the 29 Spanish phones in total) of a segment of an EMG signal using a neural network is 50%, when using data from one session and one speaker (Salomons et al., 2023). A problem with this technological approach is that the machine learning model is usually a black box. Consequently, it is unclear how each signal contributes to the recognition of certain speech units. To gain more insight into this, we attempt to use generalized additive modeling (GAM; Wood (2017)) to assess whether this technique can be used to identify the role of different facial muscles in speech production. To use GAMs most effectively, we compare the EMG signal's RMS values between a pair of words that are partially overlapping in their pronunciation. Our hypothesis is that if there is a significant difference in the EMG signal (which also takes into account the speaker-specific variability), it would be reflected in the part that also differs in pronunciation.

Methods.
As a first exploration into the potential use of GAMs, we focused on the pronunciation of a single pair of words. Specifically, from the database of Hernáez et al. (2022), we extracted the EMG signals from the minimal word pair *leche* [letʃe] and *noche* [notʃe]. Each word was repeated three to seven times by six typical speakers. For each word repetition, eight EMG signals are available, corresponding to eight superficial muscles in the face and neck: *anterior belly of the digastric* (ABD), *depressor anguli oris* (DAO), *depressor labii inferioris* (DLI), *levator labii superioris* (LLS), *masseter* (MAS), *risorius* (RIS), *stylohyoid* (SLH), and *zygomaticus major* (ZYG). For an overview of the recording setup and procedure see Salomons et al. (2023). After normalizing the time for each word, we calculated the RMS values for each of the eight signals with a window size of 25 ms and window shift of 5 ms. Then, we fitted a GAM model (following Wieling (2018)) for each muscle and visualized and assessed the significance of the difference in RMS values.

Results.
When comparing the RMS values of *noche* and *leche*, we found a significant difference for two muscles, namely the LLS and DAO (see Figure 1). The difference in muscle activation occurs in the initial part of the pronunciation, as expected.



(a) Activation pattern of the *levator labii superioris* (LLS) muscle. The LLS muscle originates from the eye socket and ends at the upper lip. Its main function is to elevate the upper lip. Time window of significant difference: 0.00 - 0.64.



(b) Activation pattern of the *depressor anguli oris* (DAO) muscle. The DAO muscle originates from the mandible and inserts into the corner of the mouth. Its main function is to depress the corner of the mouth. Time window of significant difference: 0.00 - 0.66.

Figure 1: Muscles with significant differences when comparing the EMG activation patterns for *noche* and *leche*.

Discussion. We have found that the results of this study are in line with our hypothesis. When assessing the difference in muscle activation when producing *noche* and *leche*, the difference occurs in the part where the pronunciation is different as well. The two muscles that show a significant difference are those involved in elevating and depressing the lips. They are more active when producing *noche* compared to *leche*, which is in line with the difference between the pronunciation of [o] (mouth is further open and lips are rounded) and [e] (mouth is less open and lips are not rounded). This means that GAMs applied to EMG data have the potential to be used as a method to identify muscle activation patterns while producing speech. Regarding the remaining six muscles that did not show a significant difference, we believe that they are not involved enough in speech production in general, or would turn out to be different in other contexts. In future work, we will perform a more detailed analysis of which muscles are activated in which contexts and which are not. Additionally, we will investigate whether GAMs are suitable to distinguish between audible and mouthed speech, and between typical speakers and those having had a laryngectomy. This may help to improve the EMG-based silent speech recognition method, by (for example) allowing a focus on the muscles most informative in distinguishing individual pronunciations.

References.

Diener, L., Vishkasougheh, M. R., & Schultz, T. (2020). CSL-EMG_Array: An Open Access Corpus for EMG-to-Speech Conversion. *Interspeech 2020*, 3745–3749. https://doi.org/10.21437/Interspeech.2020-2859

Gaddy, D., & Klein, D. (2020, October). Digital Voicing of Silent Speech. https://doi.org/10.48550/arXiv.2010.02960

Gonzalez-Lopez, J. A., Gomez-Alanis, A., Martin Donas, J. M., Perez-Cordoba, J. L., & Gomez, A. M. (2020). Silent Speech Interfaces for Speech Restoration: A Review. *IEEE Access*, *8*, 177995–178021. https://doi.org/10.1109/ACCESS.2020.3026579

Hernáez, I., Gonzalez Lopez, J. A., Navas, E., Pérez Córdoba, J. L., Saratxaga, I., Olivares, G., Sanchez de la Fuente, J., Galdón, A., Garcia, V., del Castillo, J., Salomons, I., & del Blanco Sierra, E. (2022). ReSSInt project: Voice restoration using Silent Speech Interfaces. *IberSPEECH 2022*, 226–230. https://doi.org/10.21437/IberSPEECH.2022-46

Salomons, I., Del Blanco, E., Navas, E., Hernáez, I., & De Zuazo, X. (2023). Frame-Based Phone Classification Using EMG Signals. *Applied Sciences*, *13*(13), 7746. https://doi.org/10.3390/app13137746

Vojtech, J. M., & Stepp, C. E. (2021, February). Electromyography. In M. J. Ball (Ed.), *Manual of Clinical Phonetics* (1st ed., pp. 248–263). Routledge. https://doi.org/10.4324/9780429320903-20

Wand, M., Janke, M., & Schultz, T. (2014). The EMG-UKA corpus for electromyographic speech processing. *Interspeech 2014*, 1593–1597. https://doi.org/10.21437/Interspeech.2014-379

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wood, S. N. (2017, May). *Generalized Additive Models: An Introduction with R, Second Edition* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Intensity downtrends in Embosi intonation

Yubin Zhang[1], Yijing Lu[1], Annie Rialland[2], Sarah Harper[3], & Louis Goldstein[1]

[1]University of Southern California

[2]Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne-Nouvelle, 4 rue des Irlandais, 75005 Paris, France

[3]Department of Neurological Surgery, University of California San Francisco, San Francisco, USA

yubinzha@usc.edu yijinglu@usc.edu annie.rialland@sorbonne-nouvelle.fr skharper@usc.edu louisgol@usc.edu

Introduction. Past research on intentional trends have focused mainly on fundamental frequency (f0) (Pierrehumbert, 1980). However, linguistic representations of intonation trends have been suggested to be much richer than f0 alone (Beckman, Hirschberg, & Shattuck-Hufnagel, 2010). The subglottal pressure and intensity variations caused by pulmonic initiatory movements, have been argued to be involved in intonational trends (Ladefoged, 1968; Lieberman, 1958; Strik & Boves, 1995). However, little is known about the dynamical properties of intensity trends in intonational contrasts and their relationship with f0 trends.

The current study testes hypotheses about trends of intensity and f0 in declarative and polar question intonations in a Bantu language called Embosi. For f0 aspects of its declarative intonation, our previous work shows that the Embosi declarative intonation exhibits initial f0 rising and final f0 hard landing. Utterance-initially, f0 rises with a positive velocity which becomes increasingly negative, while utterance-finally f0 lands hard with increasingly negative velocity. In the current study, we extend our previous work by examining the intensity and f0 trends in its declarative versus polar question intonation.

We examine three alternative task dynamics hypotheses. The pulmonic pressure initiation hypothesis states that the task variable involved in intonational trends is a pulmonic initiatory component governed by parameters that vary between declarative and question intonations. This hypothesis predicts that the variation in pulmonic initiatory movement (or subglottal pressure) causes parallel acoustic changes in both f0 and intensity. F0 and intensity are expected to have larger initial height, larger initial velocity, and smaller acceleration in polar questions than in declaratives. Alternatively, the pitch synergy hypothesis states that f0 is the task variable for global intonational trend. Pulmonic initiatory movements, together with laryngeal gestures, is part of the synergy for achieving the f0 goals. This hypothesis also predicts parallel f0 and intensity trends in declarative versus question intonation. However, it is also likely that intensity trends do not always parallel f0 trends, the independent task hypothesis states that intensity and f0 are task variables of independent global dynamical systems. This hypothesis predicts dissociation between properties of f0 and intensity trends.

Methods. The audio files for the acoustic analysis were taken from two Embosi corpora. In total, we analyzed 204 declarative and 60 polar question utterances. The mean f0 and intensity values of each mora in these utterances were extracted. To characterize the utterance-initial and utterance-final patterns of f0 and intensity trends, we fit separate linear mixed-effects models for each acoustic measure (first three moras for initial events; last three moras for final events).

Results. For f0 trends (Figure 1), the acoustic results show f0 initial rising and final lowering in both declaratives and polar questions. Utterance-initially, the polar question intonation has larger initial f0 height and initial f0 velocity than the declarative ones. Utterance-finally, the f0 difference between the two intonations is decreased. Nevertheless, the f0 in the polar question lands less hard than that in the declaratives. For intensity trends, we found evidence for initial intensity rising and final intensity hard landing, which resembles f0 trends. As with f0 trends, the question intonation has larger initial intensity than the declarative one and the intensity difference between the two utterances is reduced utterance-finally. However, intensity and f0 trends do not always match. First, there is no evidence for larger initial intensity velocity in question intonation. Moreover, f0 lands less hard in polar questions than declaratives, but intensity lands harder in polar questions.

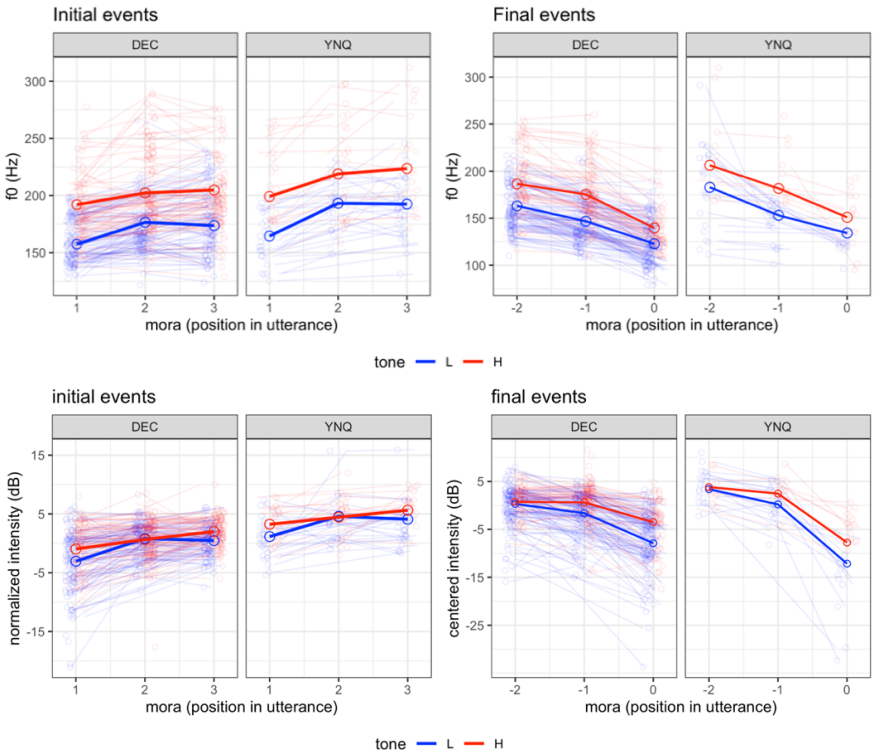


Figure 1. F0 trends and intensity trends (DEC: declarative; YNQ: polar question)

Discussion. A pure pulmonic pressure initiation hypothesis or pitch synergy hypothesis is not supported by these findings. The current evidence for dissociation between f0 and intensity patterns is more consistent with the independent task hypothesis. We propose a pulmonic pressure initiation dynamical unit in addition to the intonational tone gestures at a global utterance level. The linguistic control variable is hypothesized to be a pulmonic pressure initiatory movement variable like lung volume decrement, governed by differential equations (Catford, 1997; Zhang, 2016). Language-specific and intonation-specific variations of subglottal pressure or intensity trends are hypothesized to arise from the parameter specification of components of the pulmonic pressure initiatory dynamics. This unit is responsible for the initial rising and final hard landing intensity trends in Embosi. The f0 variations are affected by both pulmonic pressure initiation and intonational tone gestures. As a result, the less hard f0 landing in question intonation may reflect a blended result of the implementation of soft-landing L boundary tone gesture and hard-landing pulmonic pressure initiatory gesture.

References

1. Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2010). The original ToBi system and the evolution of the ToBi framework. In Prosodic Typology: The Phonology of Intonation and Phrasing. 2. Catford, J. C. (1997). Fundamental problems in phonetics. Edinburgh University Press. 3. Ladefoged, P. (1968). Linguistic aspects of respiratory phenomena. Annals of the New York Academy of Sciences, 155(1), 141–151. 4. Lieberman, P. (1958). Intonation, perception, and language. PhD dissertation, MIT. 5. Pierrehumbert, J. (1980). The phonology and phonetics of English intonation. 6. Strik, H., & Boves, L. (1995). Downtrend in F0 and Psb. Journal of Phonetics, 23(1–2), 203–220. https://doi.org/10.1016/S0095-4470(95)80043-3 7. Zhang, Z. (2016). Respiratory laryngeal coordination in airflow conservation and reduction of respiratory effort of phonation. Journal of Voice, 30(6).

Fundamental frequency as an acoustic cue to phonological phrase boundary in Spanish

Mario Casado-Mancebo[1]

[1]Universidad Nacional de Educación a Distancia

mcasado@flog.uned.es

Introduction. Prosodic structure is encoded by speakers through a range of different articulatory and acoustic cues, i.e., gestural stretching, time lengthening (Cho, 2011; Cho, 2016) and pitch movements (Baek, 2019) among other possibilities. In Spanish, Lahoz-Bengoechea (2015) confirmed the presence of prosodic cues to phonological word boundaries from a production perspective. Polo-Cano & Elordieta (2016) approached the influence of phonological phrase boundary in phonological operations that involve segmental changes. However, the actual acoustic cues to phonological phrases in Spanish are yet to be explored. This work is focused on fundamental frequency as one of these cues and so it aims to study how phonological phrase boundaries model F0 contours.

Methods. 30 Spanish speakers from Madrid were recorded reading the experiment corpus aloud in a recording booth. These texts gathered 60 sentences grouped in couples. Each couple of sentences had the same two syllables before and after a phonological word boundary (PW) or a phonological phrase boundary (PP). This is an example of a possible couple in the corpus: (a) El algo*dón de*cente no causa esos problemas, (b) Las débiles fibras de ese algo*dón de*jarán bolas al lavarlo (*Quality cotton doesn't cause those issues*, *That cotton's weak fibers will bobble after washing it*). The two sentences contain syllables *dón* and *de*. On the one hand, (a) presents these syllables in a context of a PW; that is, between a noun and its adjective (Polo-Cano, 2015; Prieto, 2006). On the other hand, (b) presents those syllables in a context of PP; that is, between a long subject and its verb (Prieto, 2006). Those recordings were then segmented using Praat TextGrids to extract the interval that spanned from the beginning of the word before the prosodic boundary to the end of the word afterwards. Python's Parselmouth (Jadoul *et al.*, 2018) was used to extract F0 data from each sample and to interpolate. Pitch floors and ceilings were adjusted manually for every single participant. To normalize duration differences between samples, a fixed number of points were extracted from the words before and after the boundary.
The whole set of F0 contours was analyzed using Functional Principal Components Analysis (Gubian *et al.*, 2015), which allows to account for factors of variation in a dataset as a function of normalized time. The number of each principal component (PC1, PC2, PC3…) expresses the decreasing amount of variance that it is able to explain (Gubian *et al.*, 2015) being PC1 the one with the largest proportion. For this study, eight PC were calculated. resulting in PC1 and PC2 being the ones of interest.

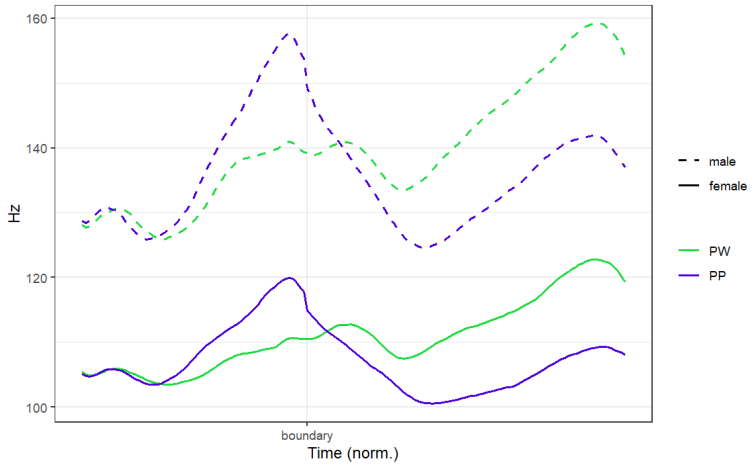


Figure 1: *Mean F0 curves.*
Males have a higher mean F0 than females and a higher rise in PP with reference to PW.

Results. PC1 and PC2 were the most informative components for this analysis. PC1 captured interspeaker variance including male/female distinction in F0. Contrary to previous literature, male participants rated a higher mean F0 than female participants (Figure 1). Two males were higher than any other participant and four women were the lowest of

all. PC2 captured prosodic boundary differences. Both PW and PP constituents cause a rise followed by a drop across the boundary. However, PP boundaries show an even higher rise and a lower drop. Although this prosodic pattern keeps between women and men, the latter spanned wider through the PC dimension. This may reflect their higher rise in PP (Figure 1). PC3 and PC4 carried some minimal adjustments to the contours and PC5-8 were non informative.

Discussion. Previous literature has established that males usually have a lower mean pitch than females. Henton (1989) presented a revision of works with references where males ranged from 68 Hz (Graddol & Swann, 1983) to 190 Hz (Philhour, 1948) and females from 126 (Graddol & Swann, 1983) to 275 Hz (Stoicheff, 1981). In the present study, however, male participants show a mean F0 ranging from 80 to 345 Hz. Pitch floor is not too far apart from the reference value, but ceiling is. Taking away from the math the two males that had extremely high F0 values results in a ceiling of 248 Hz, which is still too far from those 190 Hz in previous works. Females show a mean F0 ranging from 70 to 180 Hz. In this case, the opposite situation can be seen. Pitch ceiling is closer to reference values and it's the floor the one that is extremely low. Other factors of variation in F0 such as the time of the day when they were recorded (Grös, 2011) and vocal stress (Caraty & Montacié, 2014) where explored and discarded.
On the other hand, findings related to PP influence on F0 contours match what previous prosodic studies have set. PPs tend to align with syntactic boundaries (Selkirk, 2011), which was the starting premise for this study's corpus design. Moreover, in Spanish there seems to be a tendency to mark PP with a boundary tone (level 2 phrase break in ToBI) (Prieto, 2006), which was the main result for PC2. Contours obtained using the reconstructing function for PC2 show a generalized trend to have a H- tone before the PP boundary.

References

Baek, H. (2019). A cross-linguistic comparison on the use of prosodic cues for ambiguity resolution. 36, 060005. https://doi.org/10.1121/2.0001094

Cho, T. (2011). Laboratory phonology. En N. C. Kula, B. Botma, & K. Nasukawa (Eds.), The Continuum Companion to Phonology (pp. 343-368). Continuum.

Cho, T. (2016). Prosodic Boundary Strengthening in the Phonetics–Prosody Interface. Language and Linguistics Compass, 10(3), 120-141. https://doi.org/10.1111/lnc3.12178

Görs, K. (2011). Von früh bis spät—Phonetische Veränderungen der Sprechstimme im Tagesverlauf [BA thesis]. Kiel University.

Graddol, D., & Swann, J. (1983). Speaking Fundamental Frequency: Some Physical and Social Correlates. Language and Speech, 26(4), 351-366. https://doi.org/10.1177/002383098302600403

Gubian, M., Torreira, F., & Boves, L. (2015). Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. Journal of Phonetics, 49, 16-40. https://doi.org/10.1016/j.wocn.2014.10.001

Henton, C. G. (1989). Fact and fiction in the description of female and male pitch. Language & Communication, 9(4), 299-311. https://doi.org/10.1016/0271-5309(89)90026-8

Jadoul, Y., Thompson, B., & Boer, B. de. (2018). Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics, 71, 1-15. https://doi.org/10.1016/j.wocn.2018.07.001

Lahoz-Bengoechea, J. M. (2015). Fonética y fonología de los fenómenos de refuerzo consonántico en el seno de las unidades léxicas en español [PhD dissertation]. Universidad Complutense de Madrid.

Nespor, M., & Vogel, I. (2007). Prosodic Phonology. With a new foreword. Mouton de Gruyter. https://doi.org/10.1515/9783110977790

Philhour, C. W. JR. (1948). An experimental study of the relationship between perception of vocal pitch in connected speech and certain measures of vocal frequency.

Polo-Cano, N. (2015). Aproximación preliminar al sintagma fonológico en español. Loquens, 2(2), e020. https://doi.org/10.3989/loquens.2015.020

Polo-Cano, N., & Elordieta, G. (2016). Evidencia segmental del sintagma fonológico en español. LEA: Lingüística española actual, 38(1), 43-67.

Prieto, P. (2006). Phonological phrasing in Spanish. En Optimality-Theoretic Studies in Spanish Phonology. John Benjamins Publishing Company.

Selkirk, E. (2011). The Syntax-Phonology Interface. En J. A. Goldsmith (Ed.), The Handbook of phonological theory (pp. 435-484). Blackwell.

Stoicheff, M. L. (1981). Speaking fundamental frequency of middle-aged females. Folia phoniatrica, 19, 167-172.

Comparison of Velum Movement in /an/-rime Words between Chengdu Variety and Standard Mandarin using rt-MRI

Sishi Liao, Phil Hoole, Jonathan Harrington

Institute for Phonetics and Speech Processing (IPS), LMU Munich
sishi.liao|hoole|jmh@phonetik.uni-muenchen.de

Introduction. A sound change could be derived from phonological development (Ohala, 2012), social interactions (Labov, 1963), language contact (Boretzky, 1991), and other factors. It has been reported that the nasal coda in /an/-rime words is lost in the Chengdu variety (CD) of Southwestern Mandarin (Liao et al., 2022, 2023). However, it is yet not clear whether this sound change is the result of language contact with Standard Mandarin (SM), since the less dominant dialect variety (Chengdu) is often more prone to be subject to sound change. In this study, we collect real-time magnetic resonance imaging (MRI) recordings on native speakers from both varieties and compare the velum movement as a function of time in the target segments between these two Mandarin varieties. The objective here is to compare the velum opening between /an, aŋ/-rimed words in CD and SM, so as to investigate whether this sound change in CD results from language contact with the dominant variety (SM) or if it is a phonetically driven sound change.

Methods. This study reports data from 4 native CD speakers (2 female) and 3 SM speakers (2 female). Each participant was recorded with the 3T MRI system in a supine position and was asked to read the carrier phrases in their respective dialect. With a Mandarin syllable structure being 'CGVN', the finals of the target words (GVN) were phonologically /(G)an/ and /(G)aŋ/, with G (glide) in /ø (null), j, w/. MR images were recorded, reconstructed with a frame rate of 50 fps, and synced with noise-suppressed audio. The velum opening signal was derived from the MR images from each vocalic interval, i.e. the vowel in the CGVN syllable structure, with vocal tract aperture algorithms (Carignan et al., 2020) in MATLAB, which were then resampled to 100 data points for each observation. A total of 718 tokens (after deleting some incorrectly produced items) were analyzed and were then put into the discrete cosine transformation (DCT). The resulting DCT coefficients, k_0 and k_1 that are proportional to the mean and linear slope respectively (Watson & Harrington, 1999), were then clustered by *k-means* with two centers. The accuracies of the clustering results (with regard to the actual rime type) were compared between speaker groups.

Results. The velum opening signals as a function of time for each speaker group are shown in the left panel in Figure 1. For the native speakers of Standard Mandarin, the velum opening during the vowel segment between /an, aŋ/ rimes are quite similar: both approaching the maximum opening at the vowel offset, showing a largely opened velum pattern; meaning the nasal coda is firm. However, for the native speakers of the Chengdu variety, the velum movement approached the maximum for /aŋ/-rime, but not for /an/- rime. The right panel of Figure 1 exhibits the original rime type (in text) and the predicted rime type generated by the *k-means* clustering algorithm (in the respective color legend) for each speaker group on the $k_0 \times k_1$ space. Red color denotes observations predicted as /an/ rimes, and black for /aŋ/ rimes. The main result was that the extent of the distinction between /an, aŋ/ differed between SM and CD: for Mandarin speakers, 106 / 157 tokens (male/female) were analyzed with accuracies of 50.94% / 42.04% which is close to chance; by contrast, among Chengdu speakers, 237 / 218 tokens (male/female) were correctly categorized with scores of 93.67% / 100.00%. Thus, the result shows a similar velum movement pattern for /an, aŋ/-rimes in SM but a clear distinction with nasal loss in /an/-rime in CD.

Discussion. These results illustrate the great difference in the velum movement in /an/-rime words between two Mandarin varieties. More specifically, and as Figure 1 shows, /a/ has about the same degree of nasalization preceding /n, ŋ/ in SM, whereas in CD there are marked differences: the vowel in /an/-rime is close to oral, meanwhile the vowel in /aŋ/-rime is almost as nasalized as in SM. Despite the fact that the Standard Mandarin being the dominant variety in China, the oralization of the vowel in CD /an/-rime (Liao et al., 2022) does not appear to be a result of language contact with the SM variety. Instead, the Chengdu variety but not Standard Mandarin is participating in a phonetically motivated sound change involving /an/ → /ãn/ → /ɛ̃n/ → /ɛ̃/ → /ɛ/ that has also been observed for other languages and varieties (Hajek & Maeda, 2000; Ohala & Busa, 1995). Example words: '班' from /pan/ to /pɛ/, '攀' from /pʰan/ to /pʰɛ/.

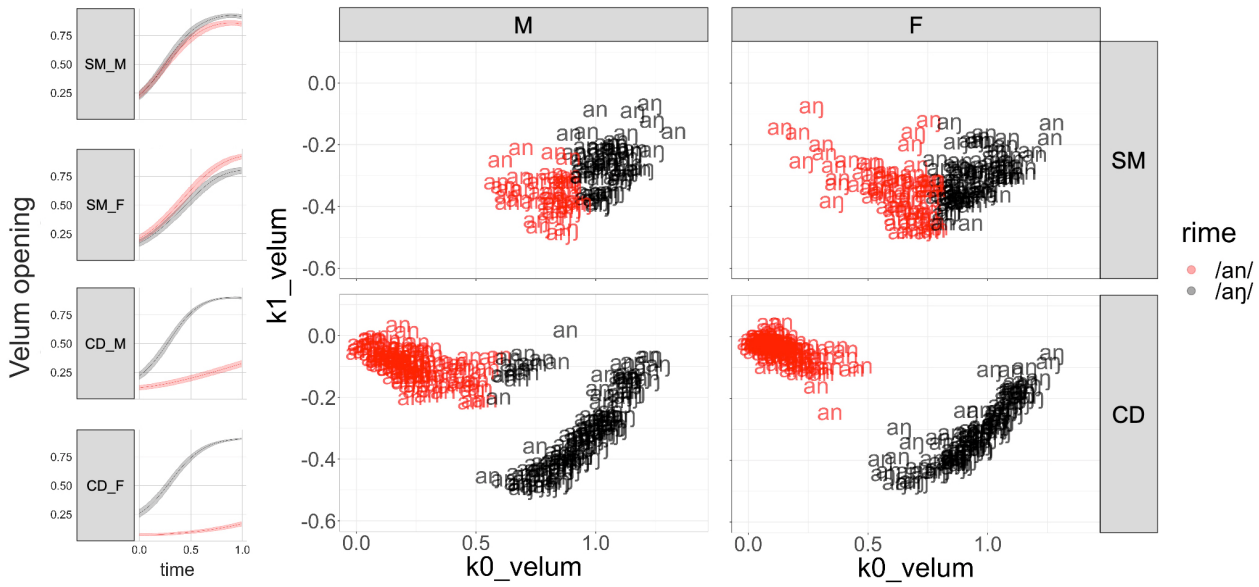


Figure 1. *The plots of the velum opening signal as a function of time for each speaker group (left panel). Each observation on the DCT coefficients $k_0 \times k_1$* space *(right panel) for each speaker group, the text of each observation denotes the actual rime type, and the color denotes the predicted type from the clustering. The color legend applies to both panels.*

References

Boretzky, N. (1991). Contact-Induced Sound Change. *Diachronica*, *8*(1). https://doi.org/10.1075/dia.8.1.02bor

Carignan, C., Hoole, P., Kunay, E., Pouplier, M., Joseph, A., Voit, D., Frahm, J., Harrington, J., Carignan, C., Hoole, P., Kunay, E., Pouplier, M., Joseph, A., Voit, D., Frahm, J., & Harrington, J. (2020). Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI. *Laboratory Phonology*, *11*(1), 1–26. https://doi.org/10.5334/LABPHON.214

Hajek, J., & Maeda, S. (2000). Investigating universals of sound change: the effect of vowel height and duration on the development of distinctive nasalization. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V* (pp. 52–69). Cambridge University Press.

Labov, W. (1963). The Social Motivation of a Sound Change. WORD, *19*(3). https://doi.org/10.1080/00437956.1963.11659799

Liao, S., Hoole, P., Cunha, C., Kunay, E., Cui, A., Shigemori, L. S. B., Kleber, F., Voit, D., Frahm, J., & Harrington, J. (2022). Nasal Coda Loss in the Chengdu Dialect of Mandarin: Evidence from RT-MRI. *Proc. Interspeech 2022*, 1347–1351.

Liao, S., Hoole, P., & Harrington, J. (2023). The relationship between vowel change and nasal loss in the chengdu dialect of mandarin Chinese: Evidence from RT-MRI. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th ICPhS* (pp. 1072–1076). Guarant International.

Ohala, J. J. (2012). The listener as a source of sound change: An update. In Maria-Josep Solé & Daniel Recasens (Eds.), *The Initiation of Sound Change* (pp. 21–36). John Benjamins.

Ohala, J. J., & Busa, M. G. (1995). Nasal loss before voiceless fricatives: a perceptually-based sound change. *Rivista Di Linguistica*, *7*, 125–144.

Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America*, *106*(1). https://doi.org/10.1121/1.427069

Production variability of Voice-Onset-Time in Spanish-speaking children

Sandy Abu El Adas[1], Marie Lallier[1]

[1]Basque Center on Cognition, Brain and Language

s.abueladas@bcbl.eu, m.lallier@bcbl.eu

Introduction. It is widely accepted that poor reading ability stem from a broad deficit in phonological awareness (Swan & Goswami, 1997). Since the main difficulty in dyslexia is in phonological processing, a considerable body of research focused on speech perception. Numerous studies have found that individuals with dyslexia show poorer identification (e.g., labeling a sound from a set of possible sounds) and discrimination (e.g., determining whether two sounds are same/different) of speech sounds (Wagner & Torgesen, 1987). However, one area in dyslexia we know little about is speech production. This is especially relevant considering the literature showing robust links between perception and production. The few studies examining production patterns in children with dyslexia have mainly used transcription-based analyses (e.g., errors, accuracy) but not acoustic-based analyses (e.g., duration, voice-onset-time) which may be more sensitive to detect differences in the speech signal (Cabbage et al., 2018). The current study addresses this gap in the literature by investigating the production variability of stops in children ages 6-7 and how production ability is modulated by reading ability.

Methods.
Participants in this study were forty-five Spanish-Basque bilingual children ages 5;0 to 5;11 (M=5;6, SD=3.41). Exposure and proficiency in each language was assessed using a parental questionnaire to ensure that all children were Spanish dominant; children that had at least 60% input and output in Spanish were included in the study. In addition, all participants met the following criteria: 1) no history of hearing and language disorders, and 2) normal verbal and nonverbal intelligence as measured by the Kbit-2 (Kaufman, 1990). To measure reading ability, each participant completed a battery reading ability which included three subtests: a word reading test, a nonword decoding test, and a letter recognition tests for both upper- and lower-case letters. In the word and nonword reading subtests children had to read words from a list of items with increasing difficulty and in the letter recognition task children were asked to name and sound the letters. A composite reading score was calculated as the average of the z-score per test for each participant. The production task consisted of twenty disyllabic (CVCV) words in Spanish with stops (p,b,t,d) in onset position and balanced for vowel context and stress pattern. To prompt repetitions, the children played an interactive game where they had to teach different aliens new words from different plants (Figure 1). The words were randomized, and tasks were elicited in multiple repetitions with breaks (20 words*5 repetitions=100 words in total, 25 productions per stop).

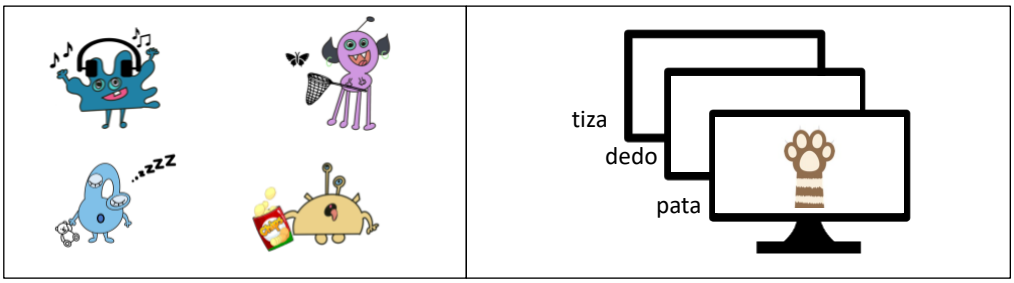


Figure 1: *right= Aliens prompting repetitions, left= example train from the production task.*

Results. Data analysis of the production data is still ongoing. Several acoustic parameters were extracted such as vowel dispersion and voice-onset-time (VOT). We extract acoustic indexes that tap into production variability, calculated by the standard deviation across the multiple repetitions. Preliminary results (N=4) of the VOT analysis revealed that children show higher rates of VOT variability for voiced stops (b,d) compared to voiceless stops (p,t). In addition, children with lower reading ability (as measured by the composite reading score) showed higher rates of VOT variability and less phonemic distinctions between voiced and voiceless stops than those that with higher reading scores (Figure 1&2).

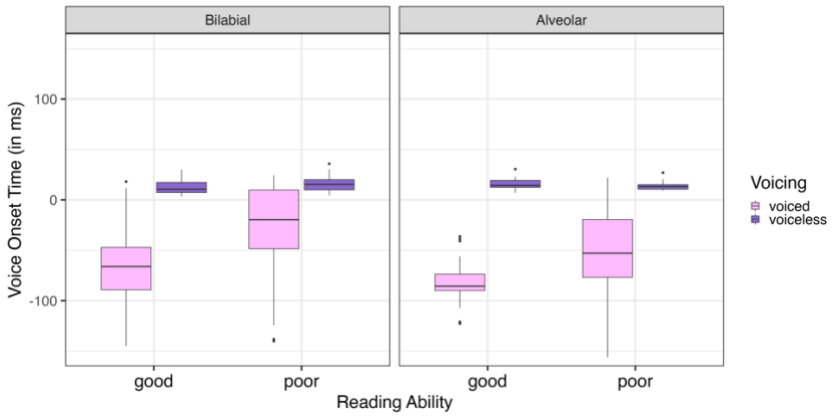


Figure 2: *Voice-Onset-Time (in ms) by place of articulation (bilabial=left, alveolar=right), voicing (voiced=pink, voiceless=purple), and reading ability (good, poor).*

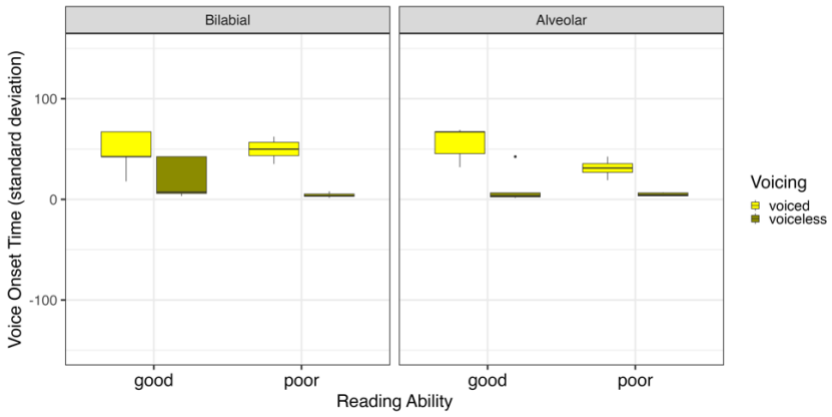


Figure 3: *Voice-Onset-Time (standard deviation) by place of articulation (bilabial=left, alveolar=right), voicing (voiced=yellow, voiceless=olive), and reading ability (good, poor).*

Discussion. Speech production variability and reading ability in children ages 5-6 was examined using a naming task and a battery of reading measures. The results from the study suggest a relationship between speech production ability and reading skills in children ages 5-6. This work provides preliminary evidence for production markers in reading difficulties and may ultimately help clinicians identify acoustic indexes associated with poor reading abilities at an earlier stage.

References

Cabbage, K. L., Farquharson, K., Iuzzini-Seigel, J., Zuk, J., & Hogan, T. P. (2018). Exploring the overlap between dyslexia and speech sound production deficits. *Language, Speech, and Hearing Services in Schools*, *49*(4), 774–786.

Kaufman, A. S. (1990). *Kaufman brief intelligence test: KBIT*. AGS, American Guidance Service Circle Pines, MN. http://www.v-psyche.com/doc/MENTAL%20ABILITY/Kaufmann%20Brief%20Intelligence%20Test%20%20second%20edition%20(KBIT-2).doc

Swan, D., & Goswami, U. (1997). Phonological awareness deficits in developmental dyslexia and the phonological representations hypothesis. *Journal of Experimental Child Psychology*, *66*(1), 18–41.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, *101*(2), 192.

Effects of an ultrasound biofeedback session on maximal tongue movements

Eija M.A. Aalto[1], Minoru Yoshida [1] Lucie Ménard[2], Walcir Cardoso[3], Catherine Laporte[1]

[1]École de technologie supérieure, [2] Université du Québec à Montréal, [3] Concordia University

Eija.aalto@etsmtl.ca, minoru.yoshida.1@ens.etsmtl.ca, menard.lucie@uqam.ca, walcir.cardoso@concordia.ca, Catherine.laporte@etsmtl.ca

Introduction. Ultrasound (US) imaging is emerging as a promising visual articulatory biofeedback device in second language (L2) learning (e.g., Chang, 2023; d'Apolito, 2017; Mozaffari & Lee, 2021), as it can assist L2 learners to visualize otherwise invisible internal articulators like the tongue and its movements. This visual feedback can aid L2 learners to approximate the unfamiliar tongue configuration required to produce the target L2 sounds, thus contributing to learning (Ouni, 2014). However, PICO (i.e. patient/population, intervention, comparison and outcomes) studies have shown that it is still unclear what kind of learners benefit the most from US biofeedback in comparison with traditional auditory-based methods (e.g. Chang, 2023; deJong, 2021; d'Apolito, 2017; Cleland & al., 2015), since individual factors may predict learning outcomes (Wong et al. 2017). For example, an individual's ability to control tongue movements is seen as a possible element affecting their performance (d'Apolito et al., 2017; Li et al., 2019). Since the ability to learn tongue movements and the benefits of US visual biofeedback are still poorly understood, our aim is to examine the effectiveness of a short biofeedback session for maximal tongue movements: tongue retraction and lowering.

Methods. The participants were six bi- or multilingual adults without any history of speech and language diagnoses. Four of the participants had no previous exposure to US biofeedback (NE), while two participants had previous exposure (PE). The participants received a short introduction to US including basic information of its function, model videos of US tongue imaging, and explanation of the articulatory movements and stuctures displayed in the videos.

The data collection session included a pre-test of maximal tongue retraction and lowering, a two minute practice using US biofeedback, and a post-test performed immediately after the practice. The participants were asked to clench their teeth to control the jaw openness and thus minimize unwanted US probe movements. They were then asked to move their tongue as far back and as low as possible within their oral cavity. During the two minute practice, the participants accessed real time US biofeedback. They were encouraged to find the maximal tongue movements.

The US data were recorded with Telemed MicrUs EXT-1H using MC4-2R20S-3 transducer. The transducer was kept stable under the participants' mandible by securing it with two elastic bands (behind the ear and at the temple) to a helmet suspender. The adequacy of the field of view was assessed by ensuring that it covered the full tongue surface when producing the syllables /ti/ and /ga/. The midsagittal tongue contours of the maximal retraction and lowering points were extracted manually. The reference points, [g] in /ga/, [t] in /ti/, and the genioglossus tendon, were marked to before and after pictures. To compare the participants' maximal tongue movements before and after the experimental session, their before and after practice tongue contour pictures were superimposed on one another and the reference points were matched to compensate for possible transducer movements.

Results. The maximal tongue retraction and lowering pre- and post- practice are presented in **Figure 1.** In the tongue retraction dimension, most of the NE participants increased their maximal tongue retraction after the practice at least to some extent. The PE participants showed almost no change. In the tongue lowering dimension, one participant in the NE group showed a clear improvement, while the remaining participants across both NE and PE groups exhibited only negligible improvement or no noticeable change.

Discussion. The current study examined whether a very short exposure to US biofeedback affects maximal tongue movements. The results in this pilot study suggest that participants without previous exposure to US biofeedback can improve their maximal tongue movements, at least to some extent, with a very short US biofeedback session. However, as a preliminary investigation with a small sample size (n=6) and no control condition, the changes observed cannot be solely attributed to effects of US biofeedback over other types of practice. Overall, the results align with those by Ouni (2014), where improvements of tongue shapes were shown with US biofeedback. The trend of improving maximal

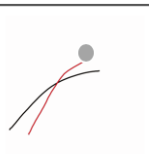
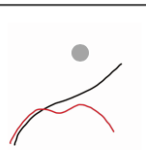
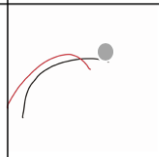
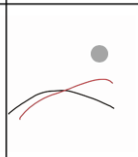
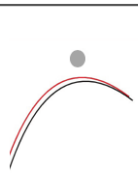
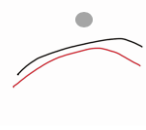
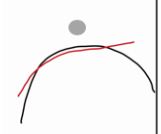
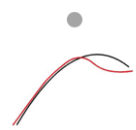
	Without previous exposure				With previous exposure	
Participant	1	2	3	4	5	6
Maximal tongue retraction						
Maximal tongue lowering						

Figure 1: Pre- (black line) and post- (red line) biofeedback practice maximal tongue movements of the participants. The gray circle shows the place of articulation of [g] in /ga/ syllable. The tongue tip is on the right.

movements seems positive, since no participants demonstrated reduced movements following the practice. Rather, all participants either exhibited no change or increased in the two movement dimensions. Furthermore, the changes were less noticeable in participants with previous exposure to biofeedback. The tongue lowering task was more challenging than the retraction while clenching the teeth for both groups, and the changes with biofeedback are modest except for participant 2, who seemed to realize how to lower the tongue with the US biofeedback.

The short biofeedback session utilized in this study could be considered as "warm up" to phoneme practice using US biofeedback, and as a measure of tongue movement awareness and ability to learn new motor movements fast, and thus provide information of individual differences that may affect US biofeedback outcomes (e.g., d'Apolito et al., 2017; Li et al., 2019). The small sample already showed greater gains in maximal movements in some participants in the NE group. These preliminary results may provide baseline articulatory performance metrics, especially when data of a few additional simple movements is collected. The participants' ability to quickly modify their tongue movements with US biofeedback will be used to inform participant group stratification for an ongoing experiment investigating US biofeedback for L2 pronunciation training.

References

Chang, Y. H. S. (2023). Effects of Production Training With Ultrasound Biofeedback on Production and Perception of Second-Language English Tense–Lax Vowel Contrasts. *Journal of Speech, Language, and Hearing Research*, *66*(5), 1479-1495.

Cleland, J., Scobbie, J. M., Nakai, S., & Wrench, A. A. (2015, August). Helping children learn non-native articulations: The implications for ultrasound-based clinical intervention. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS), Glasgow, 10-14 August 2015*. International Phonetic Association.

d'Apolito, I. S., Sisinni, B., Grimaldi, M., & Fivela, B. G. (2017). Perceptual and ultrasound articulatory training effects on English L2 vowels production by Italian learners. *International Journal of Cognitive and Language Sciences*, *11*(8), 2174-2181.

de Jong, L., Rebernik, T., Vaziri, S., & Wieling, M. (2021). Using ultrasound tongue imaging to improve L2 English pronunciation in Dutch students. In *12th International Seminar on Speech Production* (pp. 60-63). Haskins Press.

Li, J. J., Ayala, S., Harel, D., Shiller, D. M., & McAllister, T. (2019). Individual predictors of response to biofeedback training for second-language production. *The Journal of the Acoustical Society of America*, *146*(6), 4625-4643.

Mozaffari, M. H., & Lee, W. S. (2021, December). Second Language Pronunciation Training by Ultrasound-enhanced Visual Augmented Reality. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 3043-3050). IEEE.

Ouni, S. (2014). Tongue control and its implication in pronunciation training. *Computer Assisted Language Learning, 27*(5), 439-453.

Wong, P. C., Vuong, L. C., & Liu, K. (2017). Personalized learning: From neurogenetics of behaviors to designing optimal language training. *Neuropsychologia*, *98*, 192-200.

The role of face and head movement in the production of lexical tones in Cantonese

João Vítor Possamai de Menezes[1], Maria Mendes Cantoni[2], Hani Camille Yehia[3], Denis Burnham[4], Adriano Vilela Barbosa[3]

[1]*Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany*
[2]*Faculty of Letters, Federal University of Minas Gerais, Brazil*
[3]*Department of Electronic Engineering, Federal University of Minas Gerais, Brazil*
[4]*MARCS Institute for Brain, Behavior & Development, Western Sydney University, Australia*
joao_vitor.possamai_de_menezes@tu-dresden.de

Introduction. Tone languages are characterized by the use of lexical or grammatical tones, which may be defined as pitch variations systematically associated with changes in the core meaning or usage of a word. It is estimated that around half of the world population speaks tone languages (Yip 2002), hence the relevance of such languages as a subject of study. The multimodality of speech, both at the production and perception ends, is investigated since the 1950s (see, for example, the seminal work of Sumby and Pollack (1954)) and motivated studies on the auditory-visual perception of speech at the segmental and, later, suprasegmental levels. In this context, the multimodality of lexical tones became a research subject. Even though lexical tones are generally characterised by pitch patterns, visual patterns such as the movement of the head as a rigid body and of the individual parts of the face also play a role in lexical tone perception (Garg et al. 2019; Menezes et al. 2020; Burnham et al. 2022). This study is motivated by previous results from our group (Menezes et al. 2020; Burnham et al. 2022) where Cantonese lexical tones could be successfully determined based solely on visual information recorded with an Optotrak device. The current work continues our previous investigations by i) adding more speakers (three instead of one) to our analysis, ii) using more face markers, compared to Burnham et al. (2022), in order to track eyebrow movement, and iii) conducting a more detailed analysis, compared to Menezes et al. (2020), of the contribution of individual face and head motion components to tone classification.

Methods. The speech production experiments for data acquisition were conducted at the MARCS Institute for Brain Behavior and Development (Sydney, Australia) with 3 native speakers of Cantonese. The corpus was composed by 216 isolated words in Cantonese, which were combinations of 36 phonetic strings with the 6 lexical tones. The corpus was recorded 4 times for each speaker, with the audio and Optotrak camera data being recorded synchronously. The recorded Optotrak data consists of the 3D position of 33 markers attached either to the speaker's face (markers 5 through 35 in Figure 1) or to a headgear worn by the speaker (markers 1 through 4), sampled at 60 Hz. Audio data was recorded at 44100 Hz. A head motion compensation procedure was applied to the recorded marker trajectories to separate them into their two underlying components, namely, the rigid body motion of the head (6D) and the movement of the face relative to the head (3D position of 29 markers). F0 contours were extracted from the recorded audio signals in Praat using the autocorrelation method. Linear Discriminant Analysis (LDA) models were trained to classify between the 6 Cantonese lexical tones based on these 3 sets of signals (F0, head motion and face motion). LDA requires all input signals to have the same dimension, which, in our case, means all recorded words should have the same duration. As this is not the case, the dimension of the input space needed to be normalized. This was done by approximating the trajectories of each signal by a 3rd order polynomial (4 coefficients), setting the length of all input tokens to the same value. The polynomial coefficient representations of F0, head motion and face motion were centered and scaled before each LDA model was trained.

Results. For each input domain (F0, head motion, face motion), classification performance was calculated as the average accuracy over 60 repetitions of 5-fold cross validated LDA models. When all 3 speakers are considered together, the F0 accuracy was $66.94\% \pm 1.67\%$, the face motion accuracy was $50.55\% \pm 2.07\%$, and the head motion accuracy was $33.85\% \pm 1.91\%$. In order to visualize how relevant different types of face and head movements were to these results,

two analyses were performed: an inspection of the LDA rotation matrix and an ablation study. The rotation matrices of 2 LDA models (one using face motion and another using head motion as input) trained to classify between level and contour tones (2 classes) were inspected. Using just 2 classes allows a greater interpretability of the LDA rotation matrix, since its dimension is given by the number of classes minus 1. Results are shown in Figure 1 where, for clarity, face markers were clustered into 5 face regions (larynx, jaw, lips, cheeks and eyebrows). The most relevant face motion component was eyebrow movement, whereas the most relevant head motion component was translation along the x-axis, followed by head pitch. In turn, the ablation study consisted of removing individual components from each input domain (face motion regions and head motion types) and, for each case, training an LDA model in order to see the impact of that component's removal on the model's classification accuracy. In the case of the face motion, the largest absolute decreases in classification accuracy happened when the larynx (7.83%) and the eyebrow (5.42%) markers were removed. On the other hand, in the case of the head motion, the largest absolute decreases happened when translation along the z-axis (4.73%) and row (2.33%) were removed. As a comparison to the LDA rotation matrix inspection, the absolute decreases in the absence of translation along the x-axis and head pitch were 1.94% and 1.28%, respectively.

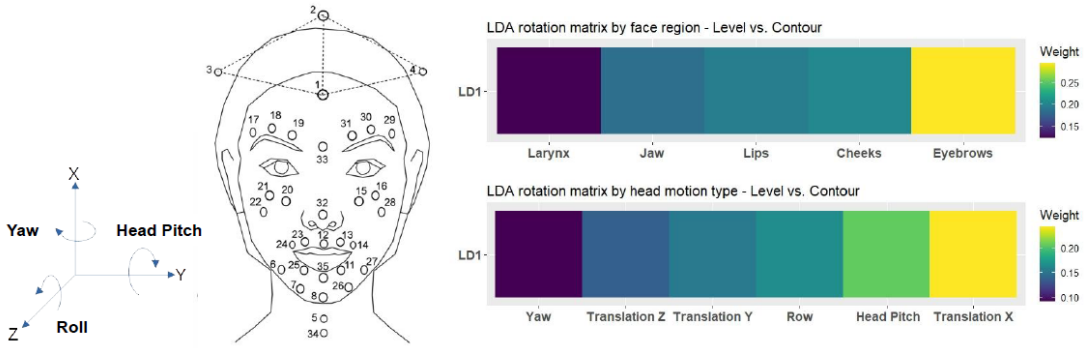


Figure 1: *Left: Axes' description. Center: Optotrak marker's positions. Right: Normalized heatmaps of face and head motion components and their weights in LDA rotation matrices differentiating Level vs. Contour tones.*

Discussion. This study has produced two main results: i) higher classification accuracy was achieved from F0 than from motion signals and ii) all motion signals were able to classify between lexical tones with above-chance accuracy. Among the investigated motion signals, higher accuracy was achieved from face motion than from head motion, confirming results in previous works. In Burnham et al. (2022), higher accuracy was achieved from head than from face motion, and this may have been due to the lack of eyebrow markers in that study. The relation between eyebrow movement and lexical tone contours suggested in Garg et al. (2019), as well as a higher accuracy obtained from face motion compared to head motion in Menezes et al. (2020) when eyebrows were included corroborate this. This study also demonstrated the importance of the eyebrows to lexical tone classification. The inspection of the LDA rotation matrices and the ablation study showed eyebrow movement as the first and second most relevant face movements, respectively. Results from the head motion analysis were not as clear. The inspection of the LDA rotation matrices indicated higher relevance of head pitch (nodding gesture, as observed in Burnham et al. (2022)) and up-down translation, whereas the ablation study indicated higher relevance of front-back translation and row (lateral rotation). Clear reasons for this were not drawn in the present study and need to be further investigated, but speaker idiosyncrasies may be at play.

References.

Burnham, Denis, Eric Vatikiotis-Bateson, Adriano Vilela Barbosa, João Vítor Menezes, Hani C. Yehia, Rua Haszard Morris, Guillaume Vignali, and Jessica Reynolds (2022). "Seeing lexical tone: Head and face motion in production and perception of Cantonese lexical tones". In: *Speech Communication* 141, pp. 40–55. DOI: https://doi.org/10.1016/j.specom.2022.03.011.

Garg, Saurabh, Ghassan Hamarneh, Allard Jongman, Joan A. Sereno, and Yue Wang (2019). "Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories". In: *Speech Communication* 113, pp. 47–62. DOI: https://doi.org/10.1016/j.specom.2019.08.003.

Menezes, João Vítor Possamai de, Maria Mendes Cantoni, Denis Burnham, and Adriano Vilela Barbosa (Sept. 2020). "A method for lexical tone classification in audio-visual speech". In: *Journal of Speech Sciences* 9.00, pp. 93–104. DOI: 10.20396/joss.v9i00.14960.

Sumby, W. H. and I. Pollack (1954). "Visual Contribution to Speech Intelligibility in Noise". In: *The Journal of the Acoustical Society of America* 26, pp. 212–215. DOI: 10.1121/1.1907309.

Yip, Moira (2002). *Tone*. Cambridge University Press. DOI: doi:10.1017/CBO9781139164559.

Schwa optionality in verbal inflection in German: the effects of stress and phonetic context

Marie-Theres Weißgerber[1]

[1]*Humboldt University Berlin*
mtweissgerber@gmail.com

Introduction. Schwa is optional in German first-person singular verbal inflectional suffixes. Variation in schwa realisations has been documented for the German language system for centuries (Fleischer et al. 2018; Nübling et al. 2013; Eisenberg 2020). While in some cases, as in inflectional paradigms used to form the past tense without ablaut, a schwa suffix is obligatory, in other cases word-final schwa can be either pronounced or omitted without yielding any semantic change. This variation is driven by a wide range of factors, from segmental and supra-segmental parameters to articulation rate (Ernestus, Hanique, and Verboom 2015; Kienast and Sendlmeier 2000) and word frequency (Pluymaekers, Ernestus, and Baayen 2006; Kohler and Rodgers 2001; Jurafsky et al. 2001).

Research on schwa in adverbs provides further insights into why such variation might occur. In a study by Fleischer et al. optionality in word-final schwas is examined in adverbs (Fleischer et al. 2018). The authors investigate *heut(e), gern(e)* and *bald(e)* in the letters of Goethe. For *heut(e)* and *gern(e)*, a highly significant impact of the following segment was found. For both adverbs, a following vowel led to significantly fewer schwa occurrences (Fleischer et al. 2018). In the case of *gern(e)*, a sonority continuum is observed: while vowels in the following segment correlate with less schwa occurrences, final schwas occur more frequently when followed by a sonorant, and slightly more often when followed by an obstruent (Fleischer et al. 2018). These results might be rooted in a preference for a balanced alternation between vowels and consonants, whereby consonantal clusters and vowel hiatus are prevented (Fleischer et al. 2018). A small number of studies found effects of different registers of spoken language on schwa realisations. Kohler and Rodgers (2001) examine schwa in both read and spontaneous speech and find that the segment articulated after a potential word-final schwa influences whether or not it is realised. They report that verbs, especially function words, often have a non-realised schwa in word-final position, particularly when preceding a vowel. Within that group, most unrealised schwas are found in function words and verb suffixes in the first person singular (Kohler and Rodgers 2001). Ernestus et al. find that the formality of a communicative situation affects the frequency and duration of prefixal schwas in Dutch, with less schwa realisations in "casually articulated speech" (Ernestus, Hanique, and Verboom 2015). Lange et al. discover differences in the frequency of schwa productions between the registers of free speech and task-based dialogue, with significantly more schwa productions in free conversation (Lange et al. 2023). Data on schwa optionality in different varieties of German is relatively scarce. To address this gap, the current study investigates two different varieties of German, German spoken in Germany (GGER) and German spoken in Namibia (NamGER), to generate new findings in this area. Wiese and Bracke find that there is a differentiation in register between standard German and Namibian German variants (Wiese and Bracke 2021). The majority of Namibian German speakers also speak at least two other languages, most commonly Afrikaans and English (Zimmer 2021). Kellermeier-Rehbein identifies the close relatedness of Afrikaans and English to German as a major facilitator for the incorporation of loan words and grammatical structures into Namibian German (Kellermeier-Rehbein 2016). Wiese and Bracke assert that the societal context in Namibia, which is characterised by multilingualism, makes the language receptive to the integration of diverse linguistic resources (Wiese and Bracke 2021).

This study investigates how schwa is distributed in spontaneous speech in two registers, formal and informal, in two varieties of German, asking under which circumstances schwa is realised in first-person singular inflectional verbal suffixes. Based on previous findings (Fleischer et al. 2018; Kohler and Rodgers 2001), it is hypothesised that schwa should be produced less frequently when the following syllable is unstressed. This effect is expected to be particularly marked when the following segment is an unstressed vowel and to be weaker for following sonorants or obstruents. It may be assumed that stimuli produced in the formal register will stay closer to the canonical form found in written productions and will therefore contain more schwa realisations.

Methods. Speech recordings were retrieved from two corpora containing two different varieties of German. Data of native speakers of German residing in Germany stem from a monolingual German subset of the RUEG corpus (Wiese, Alexiadou, et al. 2021). Data of speakers of Namibian German were extracted from the DNAM corpus (Zimmer et al. 2020). Participants were presented with visual material, either in the form of a video or a photograph story, of an accident. After viewing the material, subjects provided two summaries of the events that had taken place. In the formal condition, GGER participants were asked to provide a witness report to a police officer in the form of a voice message. In the informal condition, subjects summarised events to a friend in a voice message (Wiese, Alexiadou, et al. 2021; Wiese 2020). NamGER speakers spoke to a German teacher, impersonated by a researcher, in the formal condition (Zimmer et al. 2020). In the informal condition, subjects provided a summary of the events to a family member or friend present during the recordings (Zimmer et al. 2020). 88 recordings of 44 speakers (20 female) are analysed. In order to ensure data comparability, two age groups were examined in each case. For RUEG the age groups are adolescents from 13 to 19 years and adults from 20 to 37 years, and for DNAM the age range is from under 21 to 40 years. A total of 218 instances of verbs in the first-person singular are analysed in this study. The average speaking time of all participants analysed here is 68 seconds. Annotations were done manually in Praat (Boersma and Weenink 2023). Factors influencing schwa realisation were tested using chi-square tests and a logistic regression analysis using the R packages **lme4** (Bates et al. 2015) and **lmerTest** (Kuznetsova, Brockhoff, and Christensen 2017).

Results. Calculated across the whole data set, schwa is realised in 26.6 % of cases. Out of all instances, merely 32 (14.7 %) are followed by a stressed syllable and 14 (6.4 %) are followed by a pause. A Pearson's Chi-squared test for the variables *stressed* and *unstressed* and their influence on schwa realisations across the German and Namibian German varieties shows that the word stress of the following syllable has a significant influence on whether or not a schwa is articulated ($\chi^2 = 12.399$, $df = 1$, $p < 0.001$). Most instances of first-person singular verbs are pronounced without schwa when the following syllable is *unstressed* and 64.3 % of pauses are preceded by a stimulus with schwa. To assess the effect of phoneme class (*phonetic context*), the data were subset into instances of verbs preceding vowels, sonorants and obstruents. A logistic regression analysis of the factors *phonetic context* and *schwa realisations* for both varieties shows that within the vowel category, 19 % of observations are preceded by a suffix with schwa ($p < 0.001$). The subset of unrealised schwas in front of vowels are distributed to 94.3 % in front of unstressed vowels, non-realised schwa in front of sonorants can be found to 92.3 % in front of unstressed sonorants, and the subset of unrealised schwas in front of obstruents are distributed to 82.8 % in front of unstressed obstruents. This result indicates a slight sonority continuum within the distribution of schwa realisations and their interaction with the following context. In the GGER data frame, 60 % of potential word-final schwas are articulated in front of stressed syllables. This is the case in only 45.5 % in the NamGER subset. Results show that verbal suffixes are produced without schwa in 63.3 % of cases in the formal condition, and in 87.8 % of cases in the informal condition across both varieties. For the factors register and schwa, a Pearson's chi-squared test shows that register has a significant influence on schwa realisations ($\chi^2 = 15.009$, $df = 1$, $p < .001$). In the formal register, NamGER verbs are pronounced with schwa in 63.4 % of cases. In the data set of GGER, schwa is realised in 63 % of stimuli in the formal condition. Yet, the proportions are different between the two varieties in the informal condition. In NamGER, schwa is realised in only 10 % and in the GGER stimuli schwa is articulated in 15 % of cases.

Discussion. In summary, the results demonstrate that the variant without schwa is the most common realisation in verbal inflectional endings in the first-person singular in both varieties (73.4 % of all stimuli, $n = 160$). The literature predicted that schwa should be realised less often when preceding a vowel. This prediction can be confirmed with the data set analysed in the present study, where the effect of following vowels on schwa realisations is statistically significant. A slight sonority continuum is observed for unstressed following segments: the least schwa realisations are found in front of unstressed vowels, followed by unstressed sonorants and culminating in unstressed obstruents. However, the continuous tendency is largely not statistically significant. The stress of the following segment has a significant influence on whether or not a schwa is realised. Comparing the two varieties, the results show that schwa productions are evenly distributed across the formal register. In the informal register, NamGER exhibits only 10 % schwa realisations compared to 15 % in GGER. This discovery is of particular interest in the light of the variety's linguistic openness identified by Wiese and Bracke (2021), and its inclination to advance internal structural phenomena of German as noted by Wiese et al. (Wiese, Simon, et al. 2014). Is schwa-zero alternation, which seems to be an inherent structural feature of German, further progressing in informal Namibian German? This study presents additional evidence that schwa optionality is not random. Further investigation is necessary to assess the impact of register and the amplification of effects in Namibian German. Future accounts may want to analyse which specific elements cause the differences in schwa realisations between formal versus informal speech.

References.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48.

Boersma, Paul and David Weenink (2023). *Praat: doing phonetics by computer*. URL: http://www.praat.org/..

Eisenberg, Peter (2020). *Grundriss der deutschen Grammatik: Das Wort*. 5th ed. Stuttgart: J. B. Metzler.

Ernestus, Mirjam, Iris Hanique, and Erik Verboom (2015). "The effect of speech situation on the occurrence of reduced word pronunciation variants". In: *Journal of Phonetics* 48, pp. 60–75.

Fleischer, Jürg, Michael Cysouw, Augustin Speyer, and Richard Wiese (2018). "Variation and its determinants: A corpus-based study of German schwa in the letters of Goethe". In: *Zeitschrift für Sprachwissenschaft* 37.1, pp. 55–81.

Jurafsky, Daniel, Allan Bell, Michelle Gregory, and William Raymond (2001). "The effect of language model probability on pronunciation reduction". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2, pp. 801–804.

Kellermeier-Rehbein, Birte (2016). "Sprache in postkolonialen Kontexten. Varietäten der deutschen Sprache in Namibia". In: *Sprache und Kolonialismus. Eine interdisziplinäre Einführung zu Sprache und Kommunikation in kolonialen Kontexten*. Ed. by Thomas Stolz, Ingo H. Warnke, and Daniel Schmidt-Brücken. Berlin, Boston: De Gruyter, pp. 213–234.

Kienast, Miriam and Walter F Sendlmeier (2000). "Acoustical analysis of spectral and temporal changes in emotional speech". In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Kohler, Klaus J and Jonathan Rodgers (2001). "Schwa deletion in German read and spontaneous speech". In: *Spontaneous German speech: Symbolic structures and gestural dynamics* 35, pp. 97–123.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen (2017). "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13, pp. 1–26.

Lange, Robert, Bianca Sell, Megumi Terada, Malte Belz, Christine Mooshammer, and Anke Lüdeling (2023). "The phonetic realisation of verbal inflection in two dialogue registers of German spontaneous speech". In: *(Submitted for publication)*.

Nübling, Damaris, Antje Dammel, Janet Duke, and Renata Szczepaniak (2013). *Historische Sprachwissenschaft des Deutschen: Eine Einführung in die Prinzipien des Sprachwandels*. 4th ed. Tübingen: Narr Francke Attempto Verlag.

Pluymaekers, Mark, Mirjam Ernestus, and R. Harald Baayen (2006). "Effects of word frequency on articulatory durations of affixes". In: *Proceedings of Interspeech*, pp. 953–956.

Wiese, Heike (2020). "Language Situations: A method for capturing variation within speakers' repertoires". In: *Methods in dialectology XVI*. Vol. 59. Frankfurt Main: Peter Lang, pp. 105–117.

Wiese, Heike, Artemis Alexiadou, et al. (2021). "RUEG Corpus". In: URL: https://zenodo.org/record/3236069#.ZEKiFS-23fY. (visited on 09/13/2023).

Wiese, Heike and Yannic Bracke (2021). "Registerdifferenzierung im Namdeutschen: Informeller und formeller Sprachgebrauch in einer vitalen Sprechergemeinschaft". In: *Kontaktvarietäten des Deutschen im Ausland*. Tübingen: Narr, pp. 273–293.

Wiese, Heike, Horst J. Simon, Marianne Zappen-Thomson, and Kathleen Schumann (2014). "Deutsch im mehrsprachigen Kontext: Beobachtungen zu lexikalisch-grammatischen Entwicklungen im Namdeutschen und im Kiezdeutschen". In: *Zeitschrift für Dialektologie und Linguistik*, pp. 274–307.

Zimmer, Christian (2021). "Siedlungsgeschichte und Varietätenkontakt". In: *Zeitschrift für Dialektologie und Linguistik 88* H. 3, pp. 324–350.

Zimmer, Christian, Heike Wiese, Horst J Simon, Marianne Zappen-Thomson, Yannic Bracke, Britta Stuhl, and Thomas Schmidt (2020). "Das Korpus Deutsch in Namibia (DNam): Eine Ressource für die Kontakt-, Variations-und Soziolinguistik". In: *Deutsche Sprache* 48.3, pp. 210–232.

Self-supervised learning of the relationships between speech units, gestures and sounds using vocal imitation

Marc-Antoine Georges, Marvin Lavechin, Jean-Luc Schwartz, Thomas Hueber

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

firstname.lastname@gipsa-lab.grenoble-inp.fr

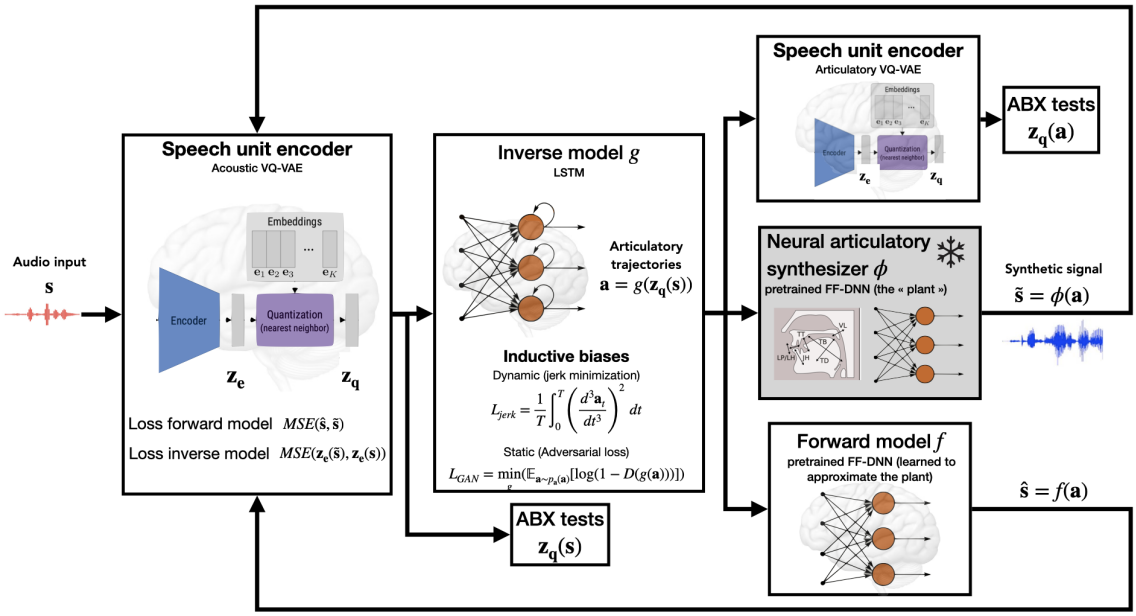


Figure 1: Overview of the proposed computational agent learning the relationships between speech units, gestures and sounds via vocal imitation.

Introduction.
Learning to speak is a hard task. It involves controlling a complex motor system for uttering speech sounds from articulatory gestures and discovering discrete and invariant speech units that enable entry into the linguistic system. Importantly, children seem to learn the relationships between speech sounds, the corresponding articulatory gestures, and these units in a weakly-supervised manner, with no explicit labeling of auditory inputs and no access to the articulatory gestures they should produce to reach an acoustic target. In this work, we study the relationships between speech units, gestures and sounds using self-supervised (deep) learning techniques. We propose a computational agent learning to drive a neural articulatory synthesizer (Georges et al. 2020) by inferring discrete speech units and articulatory commands from an auditory speech input. We evaluate the performance both at the acoustic and articulatory levels, and quantify the impact of different mechanisms (inductive biases) to regularize the ill-posed acoustic-to-articulatory inversion problem.

Architecture of the proposed agent.
The architecture of the proposed agent is presented in Figure 1.
The neural articulatory synthesizer (the plant) is a feed-forward DNN converting a vector of articulatory parameters

into a vector of ceptral coefficients. It is trained on the acoustic-articulatory French dataset PB2007[1], for which we build an articulatory model from the raw EMA data encoding 6 degrees of freedom of the vocal tract. Importantly, this neural articulatory synthesizer is used here as a pre-trained module and its parameters are not updated during the agent's training. The predicted cepstrum, combined with two source parameters (the f0 and a periodicity feature), is converted into a time-domain signal using the LPCNet neural vocoder (Valin and Skoglund 2019).

The forward internal model f predicts the acoustic consequences $\tilde{\mathbf{s}} = f(\mathbf{a})$ of the execution of a sequence of articulatory commands $\mathbf{a}$. Similarly to the neural articulatory synthesizer, it is implemented as a feed-forward DNN. However, unlike the neural articulatory synthesizer which is kept frozen all along the learning phase, the parameters of the forward model are randomly initialized. Hence the forward model contains no prior knowledge of the properties of the agent's vocal apparatus, and it must be trained by "listening" to the outputs of the agent's plant in order to learn to provide good estimates of the acoustic result of articulatory commands.

The inverse internal model g estimates the sequence of articulatory feature vectors $\mathbf{a} = [\mathbf{a}_1, \cdots, \mathbf{a}_T]$ to be sent to the synthesizer in order to approximate an auditory input $\mathbf{s} = [\mathbf{s}_1, \cdots, \mathbf{s}_T]$. It is implemented as a unidirectional recurrent neural network (LSTM).

The inductive biases aims at regularizing the ill-posed acoustic-to-articulatory inverse mapping and are additional criteria that the agent needs to optimize during the learning phase. The first one (static) prevents the inverse model from inferring out-of-domain targets, such as implausible vocal tract configurations (e.g., parts of the tongue above the palate). It takes the form of an adversarial loss that pits the inverse model against a discriminator which is trained to determine, given a training set, whether an articulatory configuration is plausible or not. The second bias encourages the agent to find an articulatory trajectory for which there is a minimum jerk from the beginning to the end of the sentence.

The speech unit encoder is a vector-quantized variational autoencoders (VQ-VAE) (Van Den Oord, Vinyals, et al. 2017) learning a codebook of quantized embeddings from the auditory input. [2]

Training algorithm: Audio input $\mathbf{s}$ is first encoded into a set of discrete embeddings $\mathbf{z_q}(\mathbf{s})$ using the VQ-VAE-based acoustic encoder. Articulatory trajectories $\mathbf{a} = g(\mathbf{z_q}(\mathbf{s}))$, inferred using the LTSM-based inverse model $\mathbf{g}$, are sent both to the pre-trained neural articulatory synthesizer (ϕ, i.e., the plant) which provides the repetition of the input stimulus by the agent $\tilde{\mathbf{s}} = \phi(\mathbf{a})$, and to the forward model f which provides the "mental" simulation of the synthesis process $\hat{\mathbf{s}} = f(\mathbf{a})$. Both internal models (forward and inverse) as well as the speech unit encoder (acoustic VQ-VAE) are jointly trained end-to-end using backpropagation. The parameters of the forward model are updated in order to minimize the mean squared error between $\widehat{\mathbf{s}}$ and $\widetilde{\mathbf{s}}$ (i.e. approximating the plant). The inverse model is then trained to minimize the acoustic reconstruction under the constrain of inductive biases, i.e., $L_{recons} = \mathit{MSE}(\mathbf{z_e}(\widetilde{\mathbf{s}}), \mathbf{z_e}(\mathbf{s})) + \lambda_g L_{GAN} + \lambda_j L_{jerk}$ (L_{GAN} and L_{jerk} being additional loss terms corresponding to the two proposed inductive biases, λ_g and λ_j are weighting factors set empirically) while keeping the forward model unchanged.

We evaluated the performance of the model both at the acoustic and articulatory levels, and quantified the impact of different mechanisms (inductive biases) to regularize the acoustic-to-articulatory inversion. ABX tests (Schatz et al. 2021) are used to assess the phonetic properties of the discrete units learned by both the speech audio encoder mentioned above (acoustic VQ-VAE), but also by an second speech unit encoder, also based on the VQ-VAE, but trained on the predicted articulatory features (articulatory VQ-VAE). Using listening tests, we showed that the model is able to "repeat" relatively complex auditory inputs[3], but has not yet succeeded in systematically producing correct articulatory trajectories. We suggest that this could be due to the lack of a realistic developmental schedule (MacNeilage 1998) likely to simplify the learning process and focus it around specific sets of commands that could provide a ground for articulatory invariance for consonantal place of articulation.

References.

Georges, M-A, P. Badin, J. Diard, L. Girin, J-L Schwartz, and T. Hueber (2020). "Towards an articulatory-driven neural vocoder for speech synthesis". In: *Proc. of ISSP (abstract)*.

MacNeilage, Peter F (1998). "The frame/content theory of evolution of speech production". In: *Behavioral and brain sciences* 21.4, pp. 499–511.

Schatz, Thomas, Naomi H Feldman, Sharon Goldwater, Xuan-Nga Cao, and Emmanuel Dupoux (2021). "Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input". In: *Proceedings of the National Academy of Sciences* 118.7, e2001844118.

Valin, Jean-Marc and Jan Skoglund (2019). "LPCNet: Improving neural speech synthesis through linear prediction". In: *Proc. of ICASSP*, pp. 5891–5895.

Van Den Oord, Aaron, Oriol Vinyals, et al. (2017). "Neural discrete representation learning". In: *Proc. of NIPS* 30.

[1] Publicly available at https://doi.org/10.5281/zenodo.6390598

[2] A vector quantized variational autoencoder (VQ-VAE) can be seen as a discrete version of a variational autoencoder for which the latent space is quantized, meaning that each latent variable is mapped to one of a set of discrete (and learned) codebook vectors

[3] Audio samples can be found at `https://georges.ma/publications/agent/`

Physiological constraints underlying the variation of labial stop intensity and spectrum

Maëva Garnier, Thibault Cattelain, Christophe Savariaux, Pascal Perrier

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
maeva.garnier@gipsa-lab.fr

Introduction. Speech intelligibility depends both on the *clarity* of different phonetically relevant cues (audible and visible), but also on their *audibility*. A speech enhancement system is potentially capable of improving all these aspects simultaneously. On the other hand, a human speaker trying to improve his intelligibility in a perturbed situation (e.g. noise) is constrained in the variation of his speech production by physical and physiological interdependencies. This raises the question as to whether it is always possible to improve clarity and audibility jointly, and to improve the clarity of all cues at once, or whether the improvement of one aspect or cue comes at the expense of another (cf. degraded clarity of shouted vowels (Rostolland & Parant, 1973)). This study aims at exploring these interdependencies in the production of labial stops, in particular in the variation of the intensity and spectrum of their "burst" (i.e. the consonant noise created at the occlusion release). Indeed, the burst spectrum is an important acoustic cue - in addition to formant transitions (Liberman et al.,1954) and other cues - for recognizing and discriminating the place of articulation of stop consonants (Stevens & Blumstein, 1978). According to theoretical models of stop consonant production (Stevens, 2000), the burst intensity is mainly determined by the level of accumulated intra-oral pressure (Badin, 1989; Hixon et al., 1967) and the articulatory velocity of occlusion release (Pelorson et al. 1997), while its spectrum is mainly determined by the cavity downstream the occlusion point, so theoretically invariant in the case of a labial stop (Stevens, 2000). Our study aims at providing in-vivo measurements of labial stop production, in order to better understand the movements by which speakers can control the variation of the intensity and spectrum of their burst, and to what extent their spectral features can be controlled "independently" of their intensity.

Methods. This study is based on the FullStop database (Cattelain 2019), in which 20 French speakers were recorded while producing repetitions of non-words /laCV/, with C={p, b} and V={a, i} in modal and whispered phonation, at comfortable and fast rates, and with increasing levels of intensity (defined subjectively by the speaker's sense of effort). The database contains the audio signal of these productions (calibrated in dB SPL), synchronized with other physiological signals, including variations in lip aperture (LA), automatically extracted from high-speed video images (200 f/s), and variations in intra-oral pressure (Pio), measured with a pneumotachograph (EVA2 system).
The burst and occlusion intervals of each production were annotated manually from the audio signal, using Praat. The average intensity of each burst was extracted, as well as three spectral descriptors: the center of gravity (CoG), skewness and kurtosis coefficients (after resampling at 8kHz and spectrum pre-emphasis). The maximum peak of Pio was measured during the occlusion phase, as well as the maximum interlip compression (from the LA signal). The lip opening velocity at occlusion release (Vop) was also measured as the maximum positive peak of the derivative of the LA signal. Multiparametric correlation analysis was conducted, at both individual and group levels, to evaluate the influence of multiple physiological parameters on the variation of each acoustic parameter (Intensity, CoG, skewness and kurtosis), depending on phonation mode, speech rate, vowel context and consonant voicing (fixed effects). Generalized linear models of the data were considered for the individual analysis, while mixed models were used for the group analysis (with the participant as a random effect).

Results. First of all, some redundancy was found between the three physiological parameters examined: for /pa/ syllables (produced in modal voice at a comfortable speech rate), the degree of interlip compression during occlusion was strongly correlated with the lip reopening velocity ($R(734)=0.81$, $p<.001$). On the contrary, it was weakly correlated with the maximum intra-oral pressure ($R(632)=0.39$, $p<.001$). In the following analyses, we therefore only considered the influence of the two physiological parameters Pio and Vop on the variation of acoustic features of the bursts.
At individual level, the burst intensity of /p/ stops (produced in modal voice, normal rate and followed by /a/) correlated significantly with both physiological parameters for 5 participants, with only one of these parameters for others (N=6 for Pio ; N=1 for Vop), or with none of them for the rest of the participants (N=8). At group level, the influence of these two physiological parameters on the burst intensity varied significantly with the phonation mode (greater influence of both Pio and Vop in whisper), with the speech rate (less influence of Pio, and loss of influence of Vop) and with the consonant voicing (greater influence of Vop, but less influence of Pio for /b/ than /p/), but not with the vowel context (cf. Figure 1). Contrary to theoretical models, our in-vivo data showed how the burst of labial stops can show a large range of spectral variations. For /p/ stops (produced in modal voice, normal rate and followed by /a/), these spectral variations were not simply predicted by variations in burst intensity (Correlation of R=-0.36, 0.27 and 0.15 for the CoG, skewness and

kurtosis, respectively). No significant correlation was found either between these spectral variations and physiological variations of Pio or Vop.

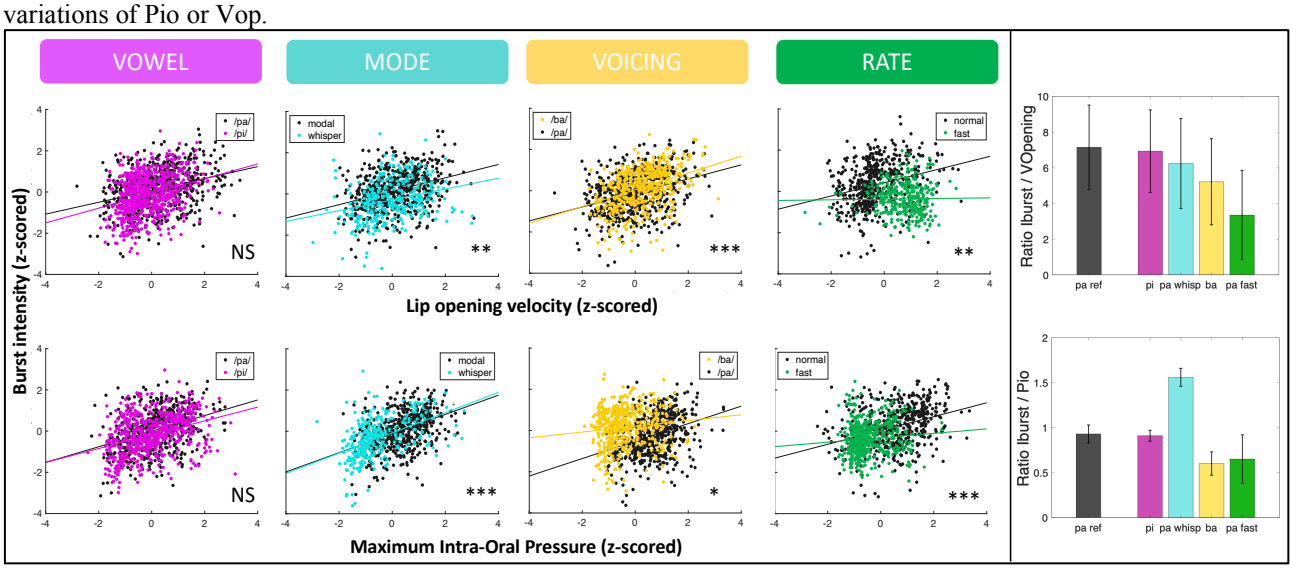


Figure 1. *Variations in burst intensity as a function of lip reopening velocity (Vop) and maximum intra-oral pressure (Pio), for /pa/ and /pi/ syllables produced in modal voice and at a comfortable rate (purple), for /pa/ syllables produced in whispered speech (turquoise blue), for /ba/ syllables (yellow) or /pa/ syllables produced at a fast rate (green). (The figures represent z-scored data by individual, to better visualize the degree of correlation between variables).*

Discussion. Our results confirm the influence of the Pio and the articulatory velocity of occlusion release, on the burst intensity of a labial stop, for more than half of the participants. However, the varying influence of these two parameters show a great variability across individuals, and an influence of phonation mode, speaking rate, and consonant voicing, suggesting the existence of varying control strategies in different individuals and situations. Furthermore, less than half the participants were able to vary the intensity of their burst relatively independently of these two physiological parameters, suggesting that there still may be other control strategies available. The high degree of correlation found between the lip compression and the lip reopening velocity suggest that both parameters reflect two aspects of a same dynamic movement of labial opening/closing. On the contrary, the weak correlation found between the lip compression and the Pio does not support the idea that lip compression is simply adjusted to contain the pressure that builds up behind the labial occlusion.

Finally, the results showed that the spectral variations of the burst of labial stops are not purely determined by variations in intensity. These spectral variations could not be explained from variations in Pio and Vop, suggesting that they might be controlled by other movements, e.g the degree of lip protrusion or the tongue position upstream the labial occlusion. Nevertheless, these results show that speakers have a certain degree of independence in controlling the intensity and spectrum of burst of labial stops, enabling them to potentially improve both the audibility and clarity of these sounds in perturbed situations of communication.

References

Badin, P. (1989). Acoustics of voiceless fricatives: Production theory and data, Speech Technol Lett, pp. 45–52.

Cattelain, T. (2019). Production des consonnes plosives du Français : du contrôle des bruit de plosion, Ph.D manuscript, Université Grenoble Alpes.

Hixon, T. J., Minifie, F. D., and Tait, C. A. (1967). Correlates of turbulent noise production for speech," J. Speech Hear. Res., 10, 133–140.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychological Monographs: General and Applied, 68(8), 1.

Pelorson, X., Hofmans, G. C. J., Ranucci, M., & Bosch, R. C. M. (1997). On the fluid mechanics of bilabial plosives. Speech Communication, 22(2-3), 155-172.

Rostolland, D., & Parant, C. (1973). Distorsion and intelligibility of shouted voice. In Proceedings of Symposium Speech Intelligibility, Liège (pp. 293-304).

Stevens, K. N. (2000). Acoustic phonetics (Vol. 30). MIT press.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. The Journal of the Acoustical Society of America, 64(5), 1358-1368.

Are glottalic mechanisms in Human Beatboxing really glottalic ?

Alexis Dehais-Underdown[1], Lise Crevier-Buchman[1,2], Didier Demolin[1],
Pierre-André Vuissoz[3], Marc Fauvel[4], Yves Laprie[5], Jacques Felblinger[3,4]

[1] *Laboratoire de Phonétique et Phonologie (CNRS/Sorbonne-Nouvelle)*

[2] *Unité Voix, Parole Déglutition, Service ORL et de Chirurgie de la Face et du Cou, Hôpital Foch*

[3] *IADI, INSERM-U947, Université de Lorraine, Nancy, France*

[4] *CIC-IT 1433, INSERM, Université de Lorraine & CHRU Nancy*

[5] *LORIA (CNRS/Inria), Université de Lorraine, Nancy, France*

alexis.dehais-underdown@sorbonne-nouvelle.fr

Introduction. Human Beatboxing is a vocal technique where artists produce musical sonorities with their vocal tract. Proctor et al. (2013) and Blaylock et al. (2017) described beatboxing using MRI data, but did not quantify the data. This paper focuses on the quantification of beatboxing MRI data on laryngeal initiation (i.e. "glottalic" initiation). A reformulation of Catford's aerodynamic model of intiation (Catford 1977) has been recently proposed (Dehais-Underdown 2023). The model was reformulated in terms of initiatory gestures. In this updated model, the so-called "glottalic" mechanisms (i.e. ejectives and implosives) are referred to as laryngeal initiatory mechanisms because they result from a sequence of laryngeal vertical movements, tongue root maneuvers and, sometimes, additional pharyngeal gestures to increase or decrease pressure in the vocal tract. Dehais-Underdown et al. (2023) also conclude tongue root maneuvers play a key role to produce glottalic mechanisms. In order to quantify laryngeal initiation and tongue root gestures, two beatboxing patterns (BP) containing ejectives and implosives produced by a single professional beatboxer were analyzed using a functional principal component analysis on selected MRI frames (n=96 frames).

Methods. *Corpus.* The corpus of this study is composed of two BP (beatboxing patterns) excerpted from a larger corpus of 11 BP. The original corpus is composed of 11 patterns with the same metrical, rhythmical and melodic structure but with different phonetic structure. Figure 1 gives an example structure of a basic pattern. The two selected patterns are [ɓ̚ ts' ʔʃ: ts' ɓ̚ ɓ̚ ts' ʔʃ: ts'] & [ɓ̚ ts' ↓ʔʃː ts' ɓ̚ ɓ̚ ts' ↓ʔʃː ts'] where [ɓ̚] and [ɗ̚] are (orally) unreleased implosives, respectively bilabial and dental; [ts'] is a dental ejective affricate and [ʔʃ:] and [↓ʔʃː] are respectively pulmonic egressive and pulmonic ingressive affricates produced with an aryepiglottal stop and a post-alveolar fricative. One professional subject repeated each pattern 4 times (4 rep. x 9 sounds x 2 BPs = 72 tokens).

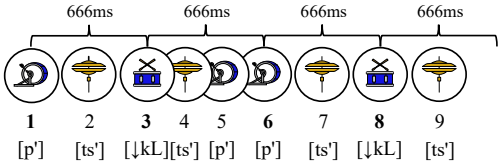
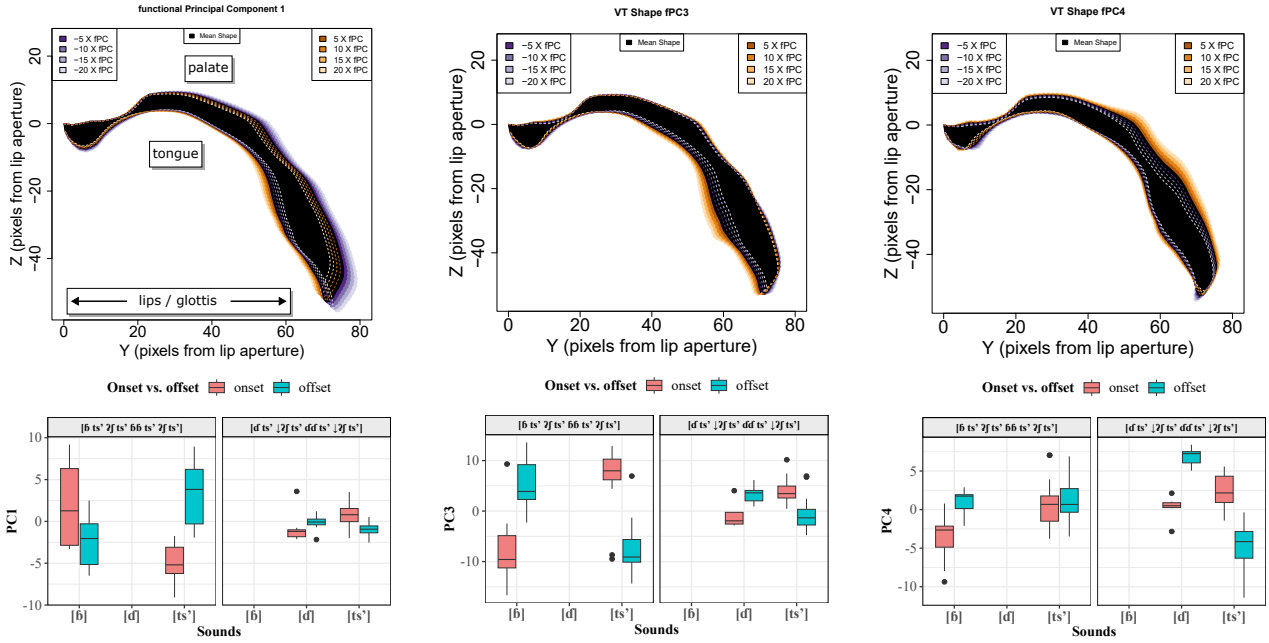


Figure 1: Metrical, rhythmical and melodic structure of BPs. Tempo of reference = 90 beat per minute

Data acquisition. MRI data of one professional beatboxer was acquired at Nancy Central Regional University Hospital with a Siemens Prisma 3T scanner, Erlangen, Germany. The subject was in supine position and a Siemens Head/Neck 64 coils was used. For the 2D real-time we used radial RF-spoiled FLASH sequence with TR = 2.22 ms, TE = 1.47 ms, FoV 192x192mm, flip angle = 5°, and slice thickness was 8 mm. Pixel bandwidth was 1670 Hz/pixel. Image size was 136x136, inplane resolution was 1.6 mm, recorded at 50 fps and reconstructed with a nonlinear inverse technique (Uecker et al. 2008). Audio was recorded at a sampling frequency of 16 kHz inside the MRI scanner with a FOMRI III optoacoustics fibre-optic microphone (FOMRI III, Optoacoustics Ltd., Mazor, Israel).
Data analysis. Semi-automatic contouring was performed on 128 selected frames using an open-source toolbox (Belyk, Carignan, and McGettigan 2023). The onset and the offset of sounds were contoured to analyze differences in vocal tract configuration at the beginning of the occlusion and at the end of the release gesture. Contours were manually corrected when needed. Contours were analyzed by means of a functional principal components analysis (fPCA) with the open source R script from Belyk and McGettigan (2022). The snares [ʔʃ:] & [↓ʔʃː] were excluded from the analysis because they are pulmonic.

Results. The five first principal components explain 87% of the variation in vocal tract configuration. fPC1 explains 37% of the variation and was found to be related to larynx height. fPC2 explains 23% of the variation but it was not clear what fPC2 was related to. fPC3 explains 15% of the variation and was found to be related to tongue root advancement and retraction. fPC4 and fPC5 each explain 6% of the variation in the data; the former was found to be related to velum movements and posterior pharyngeal wall movements and the later was found to be related to tongue height. Our analysis will focus on fPC1 (larynx height, Fig. 2a), fPC3 (tongue root movements, Fig. 2b) and fPC4 (velum height and posterior



(a) First principal component (b) Third principal component (c) Fourth principal component

Figure 2: Top : Tract variation for fPC1, 3 and 4, black contours illustrate the mean shape, orange traces indicate tract changes when a fPC increases while purple traces indicates changes when a fPC decreases. Bottom : boxplot (n=96 frames) illustrating fPC changes for each sound, colors indicate fPC variation between the onset (occlusion) and the offset (release).

pharyngeal movements, Fig. 2c). Both [ɓˀ] and [ɗˀ] are principally characterized by systematic tongue root advancement at the offset (PC3>0). For [ɓˀ] larynx height is variable at the onset and tends to be lower at the offset (PC1<0) while for [ɗˀ] the larynx is slightly higher at the offset compared to the onset. This is does not fit with the traditional view that implosives are produced by laryngeal lowering. In both cases, there is a participation of the velo-pharynx to expand the volume (closed velo-pharyngeal port and raised velum, larger pharyngeal cavity; PC4>0). [ts'] shows different behavior depending on the pattern, though tongue root retraction (PC3<0) is systematically observed at the offset. In the first BP, [ts'] is produced by raising the larynx and retracting the tongue root. In the 2nd BP, it is produced by tongue root retraction and velo-pharyngeal narrowing (PC4<0). Note that larynx is lower at the offset which does not fit with the traditional view of laryngeal raising during ejectives.

Discussion. Our data suggests that larynx vertical movements are not systematic and in some cases do not fit to the traditional view on glottalic mechanisms. Conversely, tongue root advancement ([ɓˀɗˀ]) and tongue root retraction ([ts']) are systematic. If the main gesture is tongue root advancement/retraction and not laryngeal lowering/raising, the phonetic status of beatboxed ejectives and implosives should be revised : ejectives would be characterized by retracted tongue root and implosives by advanced tongue root. If the same mechanism is found in the world's languages, then it would have important implications for sound change, for example the relationship between implosives and ATR vowels or between ejectives and pharyngeal articulations.

References.

Belyk, Michel, Christopher Carignan, and Carolyn McGettigan (2023). "An open-source toolbox for measuring vocal tract shape from real-time magnetic resonance images". In: *Behavior Research Methods*. DOI: 10.3758/s13428-023-02171-9.

Belyk, Michel and Carolyn McGettigan (2022). "Real-time magnetic resonance imaging reveals distinct vocal tract configurations during spontaneous and volitional laughter". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1863, p. 20210511. DOI: 10.1098/rstb.2021.0511.

Blaylock, Reed, Nimisha Patil, Timothy Greer, and Shrikanth S. Narayanan (2017). "Sounds of the Human Vocal Tract". In: *Interspeech 2017*. Interspeech 2017. ISCA, pp. 2287–2291. DOI: 10.21437/Interspeech.2017-1631.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Bloomington: Indiana University Press. 278 pp.

Dehais-Underdown, Alexis (2023). "Étude phonétique de la production du Human Beatbox : Approche articulatoire, aérodynamique et acoustique". Doctoral Dissertation. Université Sorbonne Nouvelle.

Dehais-Underdown, Alexis, Paul Vignes, Lise Crevier-Buchman, Didier Demolin, Pierre-André Vuissoz, Karyna Isaieva, Marc Fauvel, Yves Laprie, and Jacques Felblinger (2023). "Non-pulmonic initiation in human beatboxing: a real-time MRI study". In: *20th International Congress of Phonetic Sciences (ICPhS 2023)*. Prague, Czech Republic.

Proctor, Michael, Erik Bresch, Dani Byrd, Krishna Nayak, and Shrikanth Narayanan (2013). "Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study". In: *The Journal of the Acoustical Society of America* 133.2, pp. 1043–1054. DOI: 10.1121/1.4773865.

Uecker, Martin, Thorsten Hohage, Kai Tobias Block, and Jens Frahm (Sept. 2008). "Image reconstruction by regularized nonlinear inversion—Joint estimation of coil sensitivities and image content". In: *Magnetic Resonance in Medicine* 60.3, pp. 674–682. DOI: 10.1002/mrm.21691.

Why do palatographic data have to be taken seriously?

Yury Makarov

University of Cambridge; Institute of Linguistics, RAS

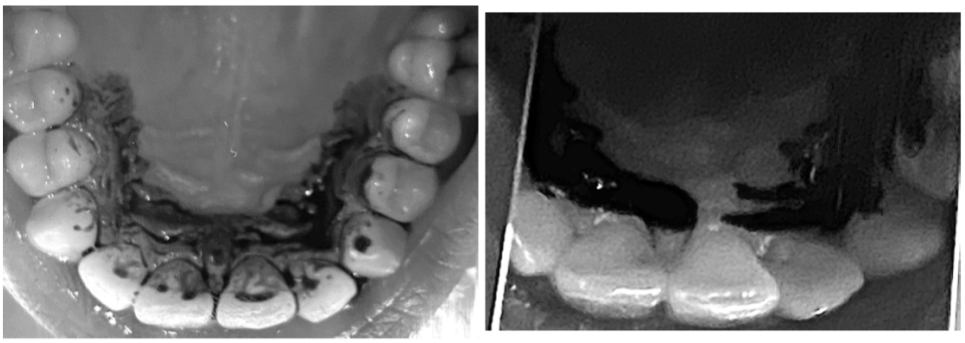
im562@cam.ac.uk

Introduction. The proposed talk concerns the use of palatography in modern-day linguistic fieldwork and provides reasons why it is essential in describing the phonetic system of any language. Furthermore, a case study of consonants of Shughni, a minor Iranian language spoken in the Pamir mountains, is discussed to show that palatography can be applied even to non-coronal articulations.

Although palatography has been known since the 19th century and does not require any special equipment except for the intraoral mirror (Ladefoged, 2003, p. 36), many linguists do not use it in their fieldwork, which is especially noticeable when it comes to the classification of coronal consonants. The attribution of coronals to dentals or alveolars often lacks any clear explanation, not to mention instrumental evidence. For example, the description of the Shughni phonemic inventory by Edelman & Dodykhudoeva (2009) states that /t d ʦ ʣ θ ð s z n r l/ are dental while Olson (2017) considers /ʦ ʣ s z n r l/ alveolar and only /t d θ ð/ are said to be dental. In both cases, no reason is given in support of either claim. The subtlety of the dental–alveolar distinction and its absence in the phonemic systems of major European languages, spoken by the scholars, may explain this discrepancy. Nevertheless, they cannot be taken as an excuse for an underworked phonetic description.

Dental–alveolar distinction. Typologically, the dental–alveolar contrast is a phonetic rarum (Molineaux, 2022, p. 663). For instance, in Urarina, an Amazonian isolate spoken in Peru, there is a distinction between the apical dental /d̪/ and apical alveolar /d/, cf. /d̪aka/ ‘wife’s brother’ vs. /daka/ ‘yesterday’ (Elias-Ulloa & Aramburú, 2021, p. 144). The contrasts of such kind tend to be marginal and unstable, and often require the support of another phonetically salient feature (Molineaux, 2022, p. 662; Wilkins, 1989, pp. 85, 88). There is a set of factors potentially influencing the dental–alveolar distinction in such phonetic systems, which includes language contacts. Provided that there are no accurate phonetic data, not only adequate phonetic/phonological descriptions (see the Shughni example above) but also the study of contact-induced phonological changes is rendered impossible.

Moreover, there is evidence that the same speaker can change their articulatory gestures associated with the same coronal phonemes. For instance, the same female speaker of Shughni, who participated in two palatographic studies in 2022 and 2023, has changed the place of articulation of /d/ and /s/ from alveolar to dental in one year, see Figure 1.



(a) 2023 (dental) (b) 2022 (alveolar; as in /ba:d/ ‘then’)

Figure 1: *Palatograms of /d/ in /ba:d/ ‘then’ for the same speaker of Shughni.*

Currently, there is no apparent factor explaining this articulatory shift; possible explanations may be learning a new language and/or physiological changes. Another problem to be considered here is allophonic or free variation within the same language. It is known that in some language varieties dental and alveolar coronals are often interchangeable realisations of the same phoneme, e.g., Scottish English as described by Wells (1982, p. 409). The study of Shughni coronals has demonstrated that seven speakers of Shughni unanimously produced /t/ and /ð/ as dentals, unlike /s/, which was alveolar in the speech of five speakers and dental in two other cases. The production of these sounds was neither influenced by the context (always the same word) nor by extralinguistic factors and can be an indication of free variation (oddly selective) or a shift from the dental articulation of /s/ to the alveolar one. The exact answer would require a series

of palatographic studies of the same language and, importantly, as many participants as possible since the variation is barely observable within two or three speakers, usually involved in palatographic research.

Shughni velars. The usability of palatograms sometimes extends beyond the realm of articulations in the front part of the mouth. The peculiar quality of velar fricatives in Shughni, characterised as 'the German *ch* of *ich* sibilated so as almost to resemble an English *sh'* by one of its first scholars (Shaw, 1877, p. 98), has attracted much linguists' attention in the 20th century. The explanations of the hissing, not typical of velars like /x/, included the grooved shape of the tongue (Sokolova, 1953, p. 137) and the raising of the tip of the tongue (Karamshoev, 1963, p. 69). Both sources, however, provided no instrumental evidence for the claims. A recent study has shown that neither of them works for the nowadays speakers of Shughni (Makarov, 2024): there are neither significant differences in the shape of the tongue compared to the typical /x/ (as in Russian) nor a sign of any front oral constriction.

Conclusion. To sum up, palatography should be included in the programme of every linguistic field research. Being quite a straightforward instrumental technique, it helps solve the basic issues of the phonemic inventory description by shedding light on the articulation of coronals. The inclusion of palatographic data into linguistic accounts has the potential to explain contact-induced changes in phonological systems and both intra- and inter-speaker articulatory variation. Palatography, despite its simplicity and lack of sophisticated laboratory equipment requirements, offers invaluable insights into the fields of articulatory phonetics, sociophonetics, and phonological typology, and should not be disregarded.

References

Edelman, D. (Joy) I., & Dodykhudoeva, L. R. (2009). Shughni. In G. Windfuhr (Ed.), *The Iranian languages* (pp. 787–824). Routledge.

Elias-Ulloa, J., & Aramburú, R. M. (2021). Upper-Chambira Urarina. *Journal of the International Phonetic Association*, *51*(1), 137–169. https://doi.org/10.1017/S0025100319000136

Karamshoev, D. (1963). *Badzhuvskij dialekt shugnanskogo jazyka [Bajuwi dialect of Shughni]*. Izdatel'stvo AN Tadzhykskoj SSR.

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Blackwell Pub.

Makarov, Y. (2024). *Shughni consonants in production: a palatographic study.* [Manuscript submitted for publication].

Molineaux, B. (2022). The dental-alveolar contrast in Mapudungun: Loss, preservation, and extension. *Linguistics Vanguard*, *8*(s5), 661–675. https://doi.org/10.1515/lingvan-2021-0080

Olson, K. (2017). *Shughni Phonology Statement*. SIL International.

Shaw, R. B. (1877). On the Shighni (Ghalchah) Dialect. *The Journal of the Asiatic Society of Bengal*, *XLVI*(2), 97–126.

Sokolova, V. S. (1953). *Ocherki po fonetike iranskikh jazykov [Outlines of the phonetics of Iranian languages]*. Izdatel'stvo Akademii Nauk SSSR.

Wells, J. C. (1982). *Accents of English: Volume 2* (Vol. 2). Cambridge University Press. https://doi.org/10.1017/CBO9780511611759

Wilkins, D. P. (1989). *Mparntwe Arrernte (Aranda): Studies in the structure and semantics of grammar* [Doctor of Philosophy]. The Australian National University.

Oral session 10

Methodology

5:30 – 6:30 pm

	Title	Authors
5:30 - 5:50 pm	Perceptual evaluation of the naturalness of broadband articulatory speech synthesis using a 1D versus a 3D acoustic model	Rémi Blandin (TU Dresden)*; Vincent Didone (University of Liège); Peter Birkholz (TU Dresden); Angélique Remacle (University of Liège)
5:50 - 6:10 pm	Advancing Speech Breathing Analysis: Benefits of Using EMA	Tabea Thies (University of Cologne)*; Philipp Buech (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne Nouvelle); Anne Hermes (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS & Sorbonne Nouvelle, Paris)
6:10 - 6:30 pm	Articulatory speech synthesis without phones	Konstantin Sering (Univeristy of Tübingen)*; Harald Baayen (University of Tübingen)

Perceptual evaluation of the naturalness of broadband articulatory speech synthesis using a 1D versus a 3D acoustic model

Rémi Blandin[1], Vincent Didone[2], Peter Birkholz[1], Angélique Remacle[3,4]

[1]*Institute of Acoustics and Speech Communication, TU Dresden, Dresden, 01062, Germany*

[2]*Psychology and Neuroscience of Cognition Research Unit (PsyNCog), Quantitative psychology, University of Liège, Liège, Belgium*

[3]*Research Unit for a Life-Course Perspective on Health and Education, Faculty of Psychology, Speech and Language Therapy, and Educational Sciences, University of Liège, Liège, Belgium*

[4]*Center For Research in Cognition and Neurosciences, Faculty of Psychological Science and Education, Université Libre de Bruxelles, Brussels, Belgium*

remi.blandin@tu-dresden.de

Introduction.

Articulatory synthesis is a useful tool to explore the relationship between the speech production and perception processes. However, including the high frequencies (HF, above about 5 kHz) requires a three-dimensional (3D) acoustical model for realistic simulations. In this frequency range, one-dimensional (1D) acoustic models fail to predict additional resonances and anti-resonances related to the 3D properties of the acoustic field. While articulatory synthesis based on 3D acoustic models is nowadays achievable for isolated phonemes, the impact of such models on the perception by human listeners remains largely unknown. Gully (2017) found that diphthongs generated with a 3D waveguide mesh were perceived as more natural than diphthongs generated with a 2D waveguide mesh and a Kelly-Lochbaum 1D model. However, the use of a time domain method reduced the quality of the simulations above 5 kHz, and the observed difference was mainly due to differences below 5 kHz. Thus, to investigate the perceptual impact of HF, a better modelling of these frequencies, and particularly of the loss mechanismes is necessary. The objective of this work was to determine whether a more realistic computation of transfer functions with a frequency domain approach results in phonemes perceived as more natural.

Methods.

Seven static phonemes, /a, e, i, ə, f, s, ʃ/, were synthesized using a source filter approach. This was done using the articulatory synthesizer VocalTractLab3D (Blandin, Arnela, et al. 2022), which can synthesize speech sounds with a 1D or a 3D acoustic model. Using geometries predefined in VocalTractLab3D, 28 stimuli were generated: 2 acoustic models (1D or 3D)×7 phonemes×2 genders. The vocal fold and aeroacoustic sound sources were simulated with a Liljencrants-Fant (LF) model, and a filtered Gaussian white noise, respectively.

Naturalness was evaluated by 31 participants aged between 21 and 28 years old (4 males and 27 females), all native French speakers. The experiment took place in a listening booth where the stimuli were played through a loudspeaker placed one meter in front of the participants. Participants listened to each stimulus as many times as they wanted and were asked to rate it on a 4 point Likert scale ranging from 0 (not at all natural) to 3 (completely natural). The stimuli were presented in a randomized order and each stimulus was rated twice at random times.

Participants' responses were analyzed with an ordinal cumulative logistic regression model using the "ordinal" R packages (Christensen 2015). The significance of the main effect (phoneme) and the interactions were assessed using a likelihood-ratio test. Contrasts (or comparisons) were made between the levels of the factors and interactions that were significant in the analysis of the models using the R packages emmeans (Lenth et al. 2019) and multcomp (Jiang and Nguyen 2007). Inter-rater reliability was assessed using the Intraclass Correlation Coefficients (ICC) (Shrout and Fleiss 1979).

Results.

Figure 1 shows the average rating for each phoneme synthesized with both acoustic models. There was no significant effect of the acoustic model (χ^2 (1) = 2.96, p = 0.085) nor the gender (χ^2 (1) = 1.13, p = 0.288). However, a significant

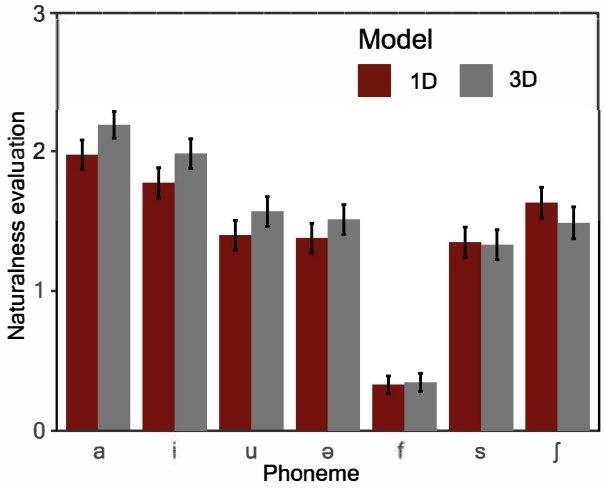


Figure 1: Average ratings for the phonemes synthesized with the 1D and 3D acoustic models in the naturalness rating task using a Likert scale from 0 (not at all natural) to 3 (completely natural).

effect of the phoneme was found (χ^2 (6) = 464, p < 0.001).
The phonemes /a/ and /i/ were rated as the most natural, with no significant difference between their ratings. /u, ə, s/ and /ʃ/ form another group with similar but lower naturalness. /f/ was rated the least natural, far below all the other phonemes, so it is mostly rated as "not at all natural".

Discussion.
In contrast to Gully (2017), our results do not show a significant influence of the 3D acoustic model on the perceived naturalness. This difference could be explained by differences in the simulation method, the phonetic material (isolated phonemes vs. diphthongs), the listening conditions (headphones vs. a loudspeaker), the experimental design (MUSHRA (Series 2014) vs. Likert scale). In fact, in Fig. 1, the average ratings are slightly higher for the 3D model of the vowels. In a subsequent study Blandin, Stone, et al. 2023, a significant difference was found using a pair comparison and a linear scale. However, only 5 vowels were used and the frequencies up to 4 kHz were similar for each model.
The highest average naturalness ratings are around 2 (rather natural), so none of the phonemes were rated as completely natural. This may be due to the material presented (isolated phonemes), geometric inaccuracies, limitations of the LF model, or remaining physical approximations (point sound source and simplified radiation).
The differences between the phonemes may be due to more or less well modeled phoneme-specific sound generation mechanisms. In particular, for fricatives: the sound source closer to the lips may be more affected by the lip shape simplification, a single point source may be too simplified for the turbulent flow aeroacoustic sound sources, their greater sensitivity to small geometric details may make them more sensitive to geometric inaccuracies, and the more directional radiation of the fricatives may be more degraded by the radiation simplifications.

References.

Blandin, R, M Arnela, S Félix, JB Doc, and P Birkholz (2022). "Efficient 3D acoustic simulation of the vocal tract by combining the multimodal method and finite elements". In: *IEEE Access* 10, pp. 69922–69938.

Blandin, R, S Stone, A Remacle, V Didone, and P Birkholz (2023). "A Comparative Study of 3D and 1D Acoustic Simulations of the Higher Frequencies of Speech". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*

Christensen, RHB (2015). *Ordinal—regression models for ordinal data, 2015. R package version 2015.6-28.*

Gully, Amelia J (2017). "Diphthong Synthesis using the Three-Dimensional Dynamic Digital Waveguide Mesh". PhD thesis. University of York.

Jiang, J and T Nguyen (2007). *Linear and generalized linear mixed models and their applications.* Vol. 1. Springer.

Lenth, R, H Singmann, J Love, P Buerkner, and M Herve (2019). "Emmeans: estimated marginal means, aka least-squares means (Version 1.3. 4)". In: *Emmeans Estim. Marg. Means Aka Least-Sq. Means https://CRAN. R-project. org/package= emmeans.*

Series, B (2014). "Method for the subjective assessment of intermediate quality level of audio systems". In: *International Telecommunication Union Radiocommunication Assembly.*

Shrout, PE and JL Fleiss (1979). "Intraclass correlations: uses in assessing rater reliability." In: *Psychological bulletin* 86.2, p. 420.

Advancing Speech Breathing Analysis: Benefits of Using EMA

Tabea Thies[1], Philipp Buech[2], Anne Hermes[2]

[1] IfL Phonetics & Department of Neurology, University Hospital Cologne, Germany
[2]Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne Nouvelle, Paris, France
tabea.thies@uni-koeln.de,{philipp.buech,anne.hermes}@sorbonne-nouvelle.fr

Introduction. The respiratory inductive plethysmograph system (RIP) is a popular technique and a validated, common tool for studying speech breathing patterns (Winkworth et al. 1995, Fuchs & Rochet-Capellan 2021, Charuau et al. 2022). Two elastic bands (with insulated wires) are positioned around the chest and the abdomen to track breathing patterns. Although different sizes of bands exist, wearing the bands may affect participants' comfort and awareness of the equipment which could further lead to alterations in breathing behavior. Another limitation is that body movements can generate artefacts in the signal that can affect the accuracy of the data (Fuchs & Rochet-Capellan 2021). Additionally, Fuchs and Rochet-Capellan (2021) pointed out that the development of smaller and/or wireless sensors could improve comfort during breathing recordings. Furthermore, to simultaneously capture kinematic speech data, one is dependent on using two recording systems, such as RIP and e.g., an electromagnetic articulograph (EMA) as has been done by e.g., Rasskazova et al. (2019). We present the use of EMA as a new applied technique for tracking speech breathing patterns, entailing high-resolution contours with better comfort and fewer artefacts. We provide the first quantitative and qualitative comparison of the RIP system (Inductotrace®) and the EMA (Carstens AG501) to track and analyze breathing patterns.

Methods. We collected acoustic and kinematic data from 18 participants (9 males, 9 females; 23-54 years, mean = 33). The kinematic breathing data were collected using the (a) EMA and (b) RIP at the same time. To track breathing data with EMA, sensors were placed at different positions and fixed with tape. One sensor on the lowest vertebra of the cervical spine functioned as a reference sensor. Sensors on the sternum and three on the chest were used to track (speech) breathing kinematics. Sensors tracking thorax movements were positioned (on clothes) at the axilla level; one in the middle and two at the height of each papilla (Fig. 1). After placing the EMA sensors, the RIP band (only chest) was put around the participants' chest. The data presented here is part of a larger data set. Here, we will present only sustained vowel productions /a/ in habitual and loud conditions to compare the similarity of the kinematic trajectories for different breathing patterns from the RIP and the EMA system. Participants were asked to take a deep breath and to produce maximum phonation of the vowel /a/ in habitual speech. In loud speech, participants were asked to keep a constant level of 80dB (distance of 1.25m of subjects to the device). The sustained /a/ phonation was repeated three times.

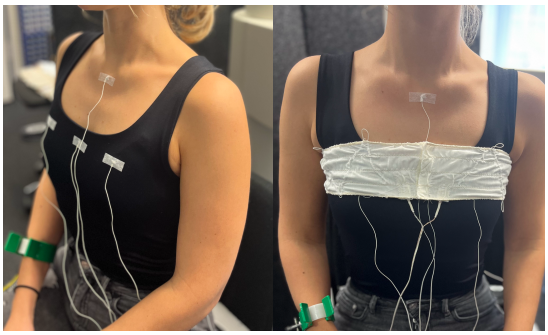


Figure 1: *EMA sensors for tracking breathing (a) before the RIP belt is put on and (b) with the RIP belt put on.*

EMA and RIP data were automatically synchronized through an impulse sent using the acoustic signal. Both EMA and RIP data were filtered (Butterworth lowpass filter, cutoff frequency 40 Hz, order 5). The distance between the reference sensor (spine) and the other sensors was analyzed in the horizontal (front-back) dimension. Three landmarks were automatically determined: (i) inhalation onset, (ii) inhalation peak, and (iii) exhalation offset (Fig. 2). Bayesian linear regression models were run on these measurements to test for the effects of the recording device. To compare the contours, we extracted the trajectories of the RIP and the EMA at 100 equally distanced time points from the inhalation onset to the exhalation offset and compared the averaged contours using a Euclidean-distance matrix (normalized by dividing the maximum distance).

Results. Here, we report the results of the distance of the EMA sensor on the chest's midpoint to the reference sensor in the low-high dimension and compare its signal with the RIP signal. At the conference, we will further provide analyses of each moving sensor and will report on which sensors are most suitable to track speech breathing with EMA. Fig. 2 provides examples for two different participants comparing the contours of the RIP and EMA (mid-chest sensor to

reference in front-back dimension) with the automatically detected landmarks. Just by visual inspection, we can see the similarity of the kinematics of both systems. Our regression analyses revealed no effect of the recording device on the inhalation and exhalation duration and their ratio. Fig. 3 displays the Euclidean-distance matrices of the two trajectories averaged across 18 speakers in habitual (Fig. 3, left) and loud conditions (Fig. 3, right). The plots show the (dis-)similarity between the two kinematic signals (across speakers and repetitions). The diagonal of each matrix represents the comparison of the trajectories at the corresponding time points. In both conditions (habitual and loud), a black diagonal beam can be observed showing a clear similarity between the trajectories of RIP and EMA.

Discussion/Conclusion. The results reveal that EMA is capable of tracking speech breathing patterns that are indeed comparable with the RIP signal. Thus, for laboratories that already own an EMA device, the use of EMA offers some practical benefits compared to the traditionally used RIP. The experimental procedure is simplified as there is no need for different belts, making it more convenient and less intrusive. In addition, EMA allows for the analysis of 3D movement patterns which are not affected by body movements due to the use of reference sensors. Furthermore, when doing EMA recordings, sensors for tracking speech breathing are easily addable to the sensor set-up when tracking articulation, making EMA a promising tool for research in speech breathing production studies. As breathing is the basic requirement for speech production and as it has a linguistic and communicative role (Fuchs & Rochet-Capellan 2021), the relevance of examining speech breathing patterns, breath cycle coordination and the interaction between breathing with other speech systems is given (Werner 2023).

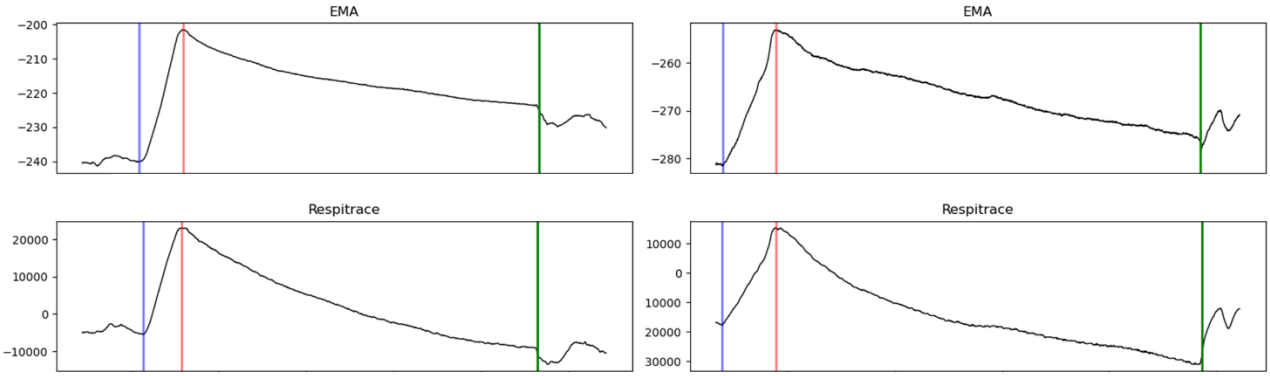


Figure 2: *Examples from (left: P07, right: P14) for habitual sustained /a/ with synchronized RIP and EMA with automatically detected landmarks: onset of inhalation (red), peak of inhalation (blue), exhalation offset (green).*

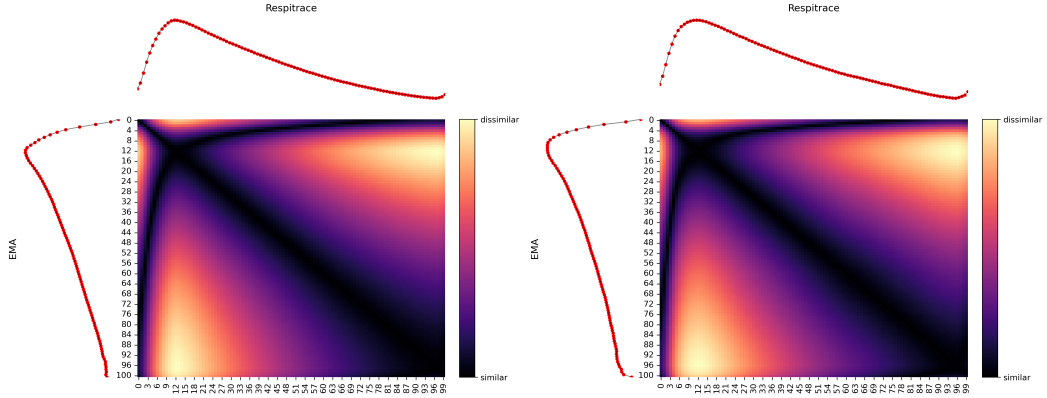


Figure 3: *Distance plots comparing RIP and EMA signals across all speakers during sustained vowel productions in habitual speech (left) and loud speech (right). Color coding: similar (black; 0 of the normalized Euclidean distance) to dissimilar (white, 1 of the normalized Euclidean distance).*

References

Charuau, D., Vaxelaire, B., & Sock, R. (2022). L'organisation spatio-temporelle de la respiration chez l'enfant. In SHS Web of Conferences (Vol. 138, p. 08005). EDP Sciences.

Fuchs, S. & Rochet-Capellan, A. (2021). The Respiratory Foundations of Spoken Language. *Annual Review of Linguistics*, 7(1), 13-30.

Rasskazova, O., Mooshammer, C. & Fuchs, S. (2019). Temporal coordination of articulatory and respiratory events prior to speech initiation. *Proceedings of 20th Interspeech* 2019, 7(1), 884-888.

Werner, Raphael Johannes (2023). The phonetics of speech breathing: pauses, physiology, acoustics, and perception. *Doctoral dissertation.* doi: 10.22028/D291-41147

Winkworth, Alison L.; Davis, Pamela J.; Adams, Roger D.; Ellis, Elizabeth (1995). Breathing Patterns During Spontaneous Speech. *Journal of Speech Language and Hearing Research,* 38(1), 124-144. doi:10.1044/jshr.3801.12

Articulatory speech synthesis without phones

Konstantin Sering[1], R. Harald Baayen[1]

[1]*University of Tübingen*

konstantin.sering@uni-tuebingen.de, harald.baayen@uni-tuebingen.de

Introduction. The Predictive Articulatory speech synthesis model Utilizing Lexical Embeddings (PAULE) is a computational model for speech production that does not use any gestures or targets on the motor side nor any phone representation on the acoustical side (Schmidt-Barbo et al. 2022). Instead it solves the task of finding suitable control parameter trajectories for the 30-dimensional speech simulator VocalTractLab (Birkholz 2013)[1] by optimizing the effect of the control in an acoustic and semantic goal space.

Several models for speech production have been proposed in the literature. Some are computationally implemented (Dell 1984; Levelt, Roelofs, and Meyer 1999), others provide more programmatic blueprints of what the production architecture might look like Fromkin (1984). What all these theories have in common is that they take sublexical units such as phonemes (the contrastive sounds of a language) and morphemes (taken to be the minimal meaning bearing units) as given, the assumption being that they provide an undisputable ground truth for theory development and computational modeling.

Another conviction shared by all these models is that production and comprehension are largely separated processes. Although, for instance, the model of Levelt, Roelofs, and Meyer (1999) takes into account that speakers are their own listeners, any systematic interaction and integration between comprehension and production is not on the horizon. In fact, the very nature of the cognitive systems underlying production and comprehension were argued by Levelt to be fundamentally different, with comprehension involving statistical inferencing from sound to phoneme sequences, but production involving a cascaded and largely interference-free sequence of selection mechanisms for lemmas, lexemes, morphemes, phonemes, and syllables.

Furthermore, the abovementioned models are static models, models that do not learn. The parameters of these models have to be set by hand. The role that experience and practice play in shaping language and language use are out of reach of these models. Finally, the cognitive models of speech production have little to say about articulation itself. The Levelt, Roelofs, and Meyer (1999) model posits that articulation is driven by syllables, which are conceived of as being, or being associated with, learned articulatory motor programs. The model by Dell (1984) likewise stops at the point that phonemes have been selected and assigned to their proper slots in phonological trees.

There are models that address articulation, but these models are found not in cognitive science, but in linguistics and phonetics. In linguistics, articulatory phonology (Browman and Goldstein 1986) posits articulatory scores. Vocal tract models, including the one implemented by VocalTractLab, create scores for control parameters by setting articulatory targets on a phoneme by phoneme basis. Smooth time series of control parameters for the different articulators are then calculated by connecting the sequences of target positions.

The PAULE model is a computational articulatory speech synthesis model that does not make any use of abstract units such as phonemes and morphemes.

Architecture. PAULE[2] implements a predictive planning approach for articulation at the word level. This predictive planning imagines the effect of the control-parameter (cp-)trajectories in terms of perceived acoustics and perceived word semantics. The cp-trajectories are smooth curves over time that define the position of the articulators as well as the parameters for the glottis model in the VTL. PAULE models all 30 control parameters of the VTL with a sampling rate of 401 Hz. For the acoustic representation a log-mel spectrogram is used with a frequency range of 10-12,000 Hz, 60 Mel bins, and a sampling rate of 200.5 Hz. For the semantic representation 300-dimensional fastText (Grave et al. 2018) vectors are used.

PAULE connects these different data structures with learned LSTM-based mappings (Hochreiter and Schmidhuber 1997). These mappings are pre-trained and back-propagate prediction errors from the semantic and acoustic representations. The

[1] https://vocaltractlab.de/index.php?page=vocaltractlab-about

[2] https://github.com/quantling/paule

back-propagated prediction error together with stationarity and constant force constraints are used to plan and optimize the control of the VTL articulatory speech synthesis model.
The LSTM-based mappings are pre-trained on a German corpus containing of 26,271 word tokens distributed over 4,311 word types. The frequency of word types follows a typical language distribution with the most common word /also/ occurring 1,113 times and 2,261 word types only occur once. The duration of the word tokens range from 120 ms to 1,000 ms. A subset of the word types, containing both long and short, and infrequent and frequent words[3], was used to evaluate PAULE.

Results.
A full implementation of the PAULE model is available for German. When given a word embedding as input, the model produces the sound waves for that word, using the VTL. The quality of the sound waves produced is sufficiently high [4] to provide (1) a strong proof of concept that a shift from mainly reactive feedforward control to predictive goal directed control is feasible and (2) that articulation without intermediate abstract sublexical units such as phonemes and morphemes is possible. Although the PAULE model currently makes use of static word embeddings, nothing prevents the use of dynamic embeddings that are specific to utterance context. Depending on the details of a dynamic embedding, the details of the articulated sound waves will change. This illustrates a more general property of the PAULE approach, namely, a shift away from what would be a 'correct' articulation to sufficiently good realizations that balance comprehensibility and minimization of articulatory effort.

Discussion.
The current implementation of PAULE has several limitations. *First*, the initialization process builds on approximate cp-trajectories synthesized from a phone-driven gesture-based approach (Sering et al. 2019). This is not a matter of principle, but a matter of convenience. Ideally, the model would be informed by either articulatory measures obtained with electromagnetic articulography or ultrasound or trained from "zero-knowledge" in a goal-babbling approach. At present, however, such empirical data are not available for the task of modeling the articulation of a non-trivial number of words. As a consequence, part of the input to the PAULE model is likely to be too systematic and rule-governed, compared to data from actual human speech. In future work, we will consider whether it is possible to obtain initialization trajectories using goal babbling learning schemes — although we anticipate that these will be computationally highly demanding. This brings us to a *second issue* we have with our model, namely, that even in its current implementation it is computationally expensive. With a realtime factor of around 3,000, planning one second of speech needs around 50 minutes of computation time. A *third issue* is that the current implementation requires several test-outs of potential articulations using the outer loop. In other words, the model is mumbling to itself before it finalizes on the articulation that it converges to as optimal. This is not how competent language users speak. Although learners need multiple try-outs to master saying a given word, mature learners have automatized what they have learned. PAULE does not utilize its memory efficiently for past experience and routinization. *Nevertheless*, we think that the PAULE model is useful as a proof of concept that considerable progress can be made in learning to articulate words using as input empirical word embeddings and the corresponding audio files within a deep learning architecture.

References.

Birkholz, Peter (2013). "Modeling consonant-vowel coarticulation for articulatory speech synthesis". In: *PloS one* 8.4, e60603.

Browman, Catherine P and Louis M Goldstein (1986). "Towards an articulatory phonology". In: *Phonology* 3, pp. 219–252.

Dell, Gary S (1984). "Representation of serial order in speech: evidence from the repeated phoneme effect in speech errors." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10.2, p. 222.

Fromkin, Victoria A (1984). *Speech errors as linguistic evidence*. Vol. 77. Walter de Gruyter.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov (2018). "Learning Word Vectors for 157 Languages". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hochreiter, Sepp and Jürgen Schmidhuber (Dec. 1997). "Long Short-term Memory". In: *Neural computation* 9, pp. 1735–80.

Levelt, Willem JM, Ardi Roelofs, and Antje S Meyer (1999). "A theory of lexical access in speech production". In: *Behavioral and brain sciences* 22.1, pp. 1–38.

Schmidt-Barbo, Paul, Sebastian Otte, Martin V. Butz, R. Harald Baayen, and Konstantin Sering (2022). "Using semantic embeddings for initiating and planning articulatory speech synthesis". In: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2022*. Ed. by Oliver Niebuhr, Malin Svensson Lundmark, and Heather Weston. TUDpress, Dresden, pp. 32–42.

Sering, Konstantin, Niels Stehwien, Yingming Gao, Martin V Butz, and Harald Baayen (2019). "Resynthesizing the geco speech corpus with vocaltractlab". In: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 95–102.

[3] Beispiel, Freunde, Lehrer, Studium, aber, eigentlich, nämlich, natürlich, praktisch, schwierig, tatsächlich, trotzdem, and zurück.
[4] Examples: `https://nc.mlcloud.uni-tuebingen.de/index.php/s/pZPgcCG9MSEHkJT`

Day 4

Friday, May 17

08:00am	
08:30am	**Jason A. Shaw**
09:00am	
09:30am	Oral Session 11
10:00am	**Coordination II**
10:30am	Coffee Break
11:00am	Poster Session 4
11:30am	
12:00am	
12:30am	
01:00pm	Lunch break
01:30pm	
02:00pm	End of the Conference

Oral session 11

Coordination II

9:30– 10:30 am

	Title	Authors
9:30 - 9:50 am	Is the articulatory trajectory of changing syllables important for achieving higher syllable rates compared to repeated syllables?	Monica Lancheros (University of Geneva)*; Marina Laganaro (University of Geneva)
9:50 - 10:10 am	Attributes Associated with Consonantal Place and Voicing in Whispered Speech	Luis Jesus (University of Aveiro)*; Sara Castilho (Hospital Arcebispo João Crisóstomo, Cantanhede); Aníbal JS Ferreira (University of Porto - Faculty of Engineering); Maria Conceição Costa (University of Aveiro)
10:10 - 10:30 am	C-G vs. C-V Timing Differences in Hong Kong Cantonese	Po-Rong Chen (National Tsing Hua University)*; Feng-fan Hsieh (National Tsing Hua University); yueh-chin chang (NTHU)

Is the articulatory trajectory of changing syllables important for achieving higher syllable rates compared to repeated syllables?

Monica Lancheros[1], *Marina Laganaro*[1]

[1]University of Geneva

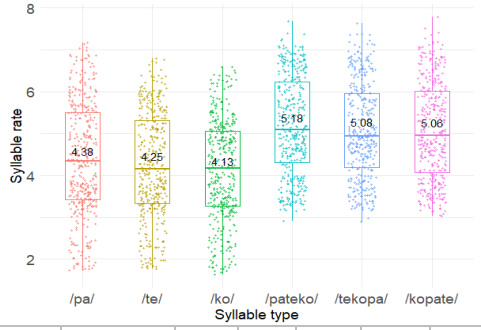
Monica.lancherospompeyo@unige.ch, marina.laganaro@unige.ch

Introduction. Speaking involves the production of complex motor sequences consisting mainly of changing articulatory movements. The production rate of syllables with varying place of articulation is commonly measured through speech-like tasks exhibiting close correspondence with speech, such as the sequential motion rate (SMR) in diadochokinetic tasks (Lancheros, Pernon, et al. 2023). They involve the repetitive production of a sequence of different syllables, such as /pataka/, at a maximum rate. Several recent studies have shown that adults achieve higher syllable rates on SMR tasks than on tasks requiring the repetitive production of single syllables (/papapa/, /tatata/) such as the alternating motion rate -AMR- tasks (e.g. Alshahwan et al. 2020; Jeng 2020; but see Lancheros, Friedrichs, et al. 2023 for different results in children and adolescents), suggesting a rate advantage of changing syllable sequences over repeated syllables. However, rate differences between SMR and AMR tasks have various unexplored interpretative hypotheses.
SMR tasks generally include syllable sequences that follow anterior to posterior articulatory trajectories, as it is the case in /pataka/ or /badaga/ (see Kent et al. 2022). One might wonder whether these movement trajectories of the oral articulators explain the ease of producing SMR sequences since the same anterior-posterior pattern is found in other oromotor acts, for instance during swallowing. Indeed, in the oral phase of this vegetative function, food is similarly propelled from the anterior to the posterior oral cavity (Logemann 2007), making this trajectory a well-rehearsed action that may provide an advantage for oral movements that follow the same pattern. In the present study, we aimed to investigate whether the high syllable rates of SMR sequences were due to the specific sequences used that involve front to back articulatory movements. For this purpose, other SMR sequences that differ from the typical labial-alveolar-velar trajectories (i.e. /pateko/) were included here: alveolar-velar-labial and velar-labial-alveolar (i.e. /tekopa/ and /kopate/, respectively). The rates of each of these SMR sequences were compared with those of AMR composed of labial, alveolar and velar phonemes (i.e. /pa/, /te/ and /ko/). If the anterior-posterior articulatory trajectory represents a real advantage for achieving higher syllable rates, then syllable rates would be expected to differ only between labial-alveolar-velar SMR sequences and AMR syllables; no differences would be expected between alveolar-velar-labial SMR and AMR or between velar-labial-alveolar SMR and AMR. However, if the articulatory trajectory of SMR sequences does not play a role in achieving higher syllable rates, then all the three SMR sequences would be expected to differ from AMR syllables.

Methods. 20 healthy French native speakers were included (5 males, mean: 23 years, range: 18–34 years). All participants gave their informed consent to participate in the study, approved by the local ethics committee, and were paid. The stimuli consisted of three AMR sequences (/papapa/, /tetete/ and /kokoko/) and three SMR sequences (/pateko/, /tekopa/ and /kopate/). The three target syllables differed in the place of articulation: bilabial with /pa/, alveolar with /te/ and velar with /ko/. They also differed in oral frequency in French: 11965.4, 14082.3 and 6101.56 occurrences per million syllables, respectively ("LEXIQUE" database, New et al. 2004). None of the SMR sequences were words in French. Regarding the procedure, participants were asked to repeatedly produce the AMR and SMR sequences as accurately as possible upon the presentation of a response cue ("?"), appearing after a short and variable delay indicated by "…". AMR and SMR targets were always presented as a 3-syllable sequence (i.e. "papapa" or "pateko", respectively). Participants were instructed to adopt a comfortable repetition rate throughout the task. They were given three seconds to repeat each sequence. Each stimulus was randomly presented 20 times each throughout the task (total = 120 items), which was divided into five blocks to allow for some breaks. Experimental runs were audio-recorded separately for each sequence. All recorded stimuli were annotated semi-automatically using the Praat software (Boersma 2001). Initially, each audio file was denoised using Praat's denoising function. Then, a script for detecting syllable nuclei (i.e. a vowel within a syllable) from intensity peaks (dB) was used to automatically count the number of syllables (De Jong & Wempe 2009). Once all syllables per participant had been detected and annotated, each audio file and its corresponding TextGrid file were checked separately to (1) determine the accuracy of the production, (2) verify the syllable count, and (3) identify the onset and offset of the production. Only correctly and fully produced syllables or syllable sequences -in the case of SMR, such as /pateko/- were included in the syllable rate calculations. Finally, the start and end time information as well as the duration of such an interval were extracted in the R software (R Development Core Team 2013) using the rPraat package (Boril & Skarnitzl 2016), from which the syllable rates were calculated. The syllable rates of the SMR sequences were compared with those of the three AMR syllables using mixed models in the R software. The base packages lmerTest and Lme4 were used with the mixed models function lmer. Contrasts were explored by using the summary function; the levels of the factor of interest were reordered using the relevel function.

Results. The mean syllable rates for the three AMR syllables and the three SMR sequences are shown in Figure 1A.

A.



B.

Syllable type	Comparisons	β	SE	df	t	p
/pateko/	/pa/	-0.80	0.14	20.02	-5.88	< 0.001
	/te/	-0.93	0.13	20.06	-6.88	< 0.001
	/ko/	-1.05	0.14	20.05	-7.43	< 0.001
/tekopa/	/pa/	-0.68	0.14	20.60	-5.04	< 0.001
	/te/	-0.81	0.13	20.74	-6.11	< 0.001
	/ko/	-0.94	0.14	20.65	-6.78	< 0.001
/kopate/	/pa/	-0.68	0.13	20.47	-5.19	< 0.001
	/te/	-0.81	0.13	20.61	-6.22	< 0.001
	/ko/	-0.94	0.14	20.55	-6.85	< 0.001

Figure 1. ***A:*** *Syllable rates (syllables per second) for each of the AMR and SMR sequences.* ***B:*** *Summary table of the mixed models contrasting SMR vs AMR sequences*

The linear mixed model (model: lmer (Rate~(syllable_type+order)+(1|item)+(1+item|participant), data = data, REML = FALSE) showed a main effect of syllable type ($F(5, 36.93) = 12.78$; $p < 0.001$) and of order ($F(1, 2310.57) = 297.98$; $p < 0.001$). Syllable type contrasts showed significant differences between the three SMR sequences and AMR syllables (see Figure 1B), suggesting that the production of alternating syllables consistently achieves higher syllable rates than the repetitive production of the same syllable, regardless of the articulatory trajectory composing the SMR sequence.

Discussion. Here we compared the performance of SMR sequences having different articulatory trajectories with that of AMR syllables to test whether the syllable rate advantage of the former sequences was due to the anterior-posterior movement of the articulators. Our results show that the three types of SMR sequences /pateko/, /tekopa/ and /kopate/ achieved significantly higher syllable rates as compared to the AMR syllables /pa/, /te/ and /ko/, suggesting that the articulatory trajectory of the SMR sequences is not responsible for changing syllables reaching higher rates compared to repetitive syllables. In a previous study we showed that the higher syllable rates found for SMR tasks are not explained by the integration of separate movement elements into ‘‘chunks’’; instead, greater anticipatory coarticulation was found in such sequences, indicating more gestural overlap across changing sequences compared to AMR syllables (Lancheros, Friedrichs, et al. 2023). Other hypotheses related to higher-level mechanisms such as motor programming inhibition for repeated syllables have been proposed (Bohland et al. 2010) but have not been investigated. Here, we also recorded electrical activity at the scalp using electroencephalography (EEG) while participants produced the AMR and SMR tasks in order to test these potential higher-level mechanisms that may play a role in preventing gestural overlap across repeated syllables (i.e., AMR tasks). The neuroimaging results will be finalized by May 2024.

References

Alshahwan, M. I., Cowell, P. E., & Whiteside, S. P. (2020). Diadochokinetic rate in Saudi and Bahraini Arabic speakers: Dialect and the influence of syllable type. *Saudi Journal of Biological Sciences*, 27(1), 303-308.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9), 341-345.

Bohland, J.W.; Bullock, D.; Guenther, F.H. Neural representations and mechanisms for the performance of simple speech sequences. J. Cogn. Neurosci. 2010, 22, 1504–1529.

Bořil, T., & Skarnitzl, R. (2016). Tools rPraat and mPraat: Interfacing phonetic analyses with signal processing. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings 19* (pp. 367-374). Springer International Publishing.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. Behavior research methods, 41(2), 385-390.

Jeng, J. Y. (2020). The vocal phonation and oral diadochokinetic movement in the elderly. *Journal of Medicine and Health*, 9(3), 41–51.

Kent, R. D., Kim, Y., & Chen, L. M. (2022). Oral and laryngeal diadochokinesis across the life span: A scoping review of methods, reference data, and clinical applications. Journal of Speech, Language, and Hearing Research, 65(2), 574-623.

Lancheros, M., Pernon, M., & Laganaro, M. (2023). Is there a continuum between speech and other oromotor tasks? evidence from motor speech disorders. *Aphasiology*, 37(5), 715-734.

Lancheros, M., Friedrichs, D., & Laganaro, M. (2023). What Do Differences between Alternating and Sequential Diadochokinetic Tasks Tell Us about the Development of Oromotor Skills? An Insight from Childhood to Adulthood. *Brain Sciences*, 13(4), 655.

Logemann, J. A. (2007). Swallowing disorders. *Best practice & research Clinical gastroenterology*, 21(4), 563-573.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3), 516-524

Attributes Associated with Consonantal Place and Voicing in Whispered Speech

Luis M. T. Jesus [1], Sara Castilho[2], Aníbal J. S. Ferreira[3], Maria Conceição Costa[4]

[1]School of Health Sciences (ESSUA), Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Intelligent Systems Associate Laboratory (LASI), University of Aveiro, Portugal
[2]Hospital Arcebispo João Crisóstomo, Cantanhede, Portugal
[3]Department of Electrical and Computer Engineering, University of Porto, Portugal
[4]Department of Mathematics (DMat) and Centre of Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal

lmtj@ua.pt, sara.castilho@ua.pt, ajf@fe.up.pt, lopescosta@ua.pt

Introduction. Whispered speech is acoustically and aerodynamically different from voiced speech (Scherer et al., 2016); it has a wider bandwidth and less peaky spectral structure, there is loss of energy at low frequencies, a flattening of high frequencies, lowering of speech rate and intensity, and lengthening of syllables or other segments, when compared to voiced speech (Meynadier, 2015; Zhang & Hansen, 2007). The sound source in whisper is a broad-band noise source generated by the exhaled air passing through a constriction, causing turbulent aperiodic airflow (Sharifzadeh et al., 2012; Sundberg et al., 2010). Whispered (phonologically) voiced consonants have been shown (Jovičić & Šarić, 2008) to be longer and have lower intensity than their voiced counterparts (reduced in intensity as much as 25 dB), but (phonologically) voiceless consonants were produced with almost unchanged intensity. Heeren (2015) found that there was no difference between voiced and whispered /f, s/ durations, their intensity was lower and the centre of gravity was lower for whispered than voiced speech. Zygis et al. (2017) showed that some spectral features of fricatives were used as segmental cues to intonation both in voiced and whispered speech. The action of the pharyngeal constrictors differs in voiced vs. voiceless pairs in both voiced and whispered speech modes (Slis & Cohen, 1969). The voiced-voiceless contrast in whispered obstruents has also been studied in various aerodynamic studies (Meynadier, 2015; Murry & Brown, 1976; Weismer & Longstreth, 1980), that have shown distinct glottal configurations and airflow volume velocity. This study explores the acoustic signal attributes that carry sufficiently distinct information to differentiate the sibilants' /s, z, ʃ, ʒ/ place and voicing in whisper. This work elaborates on a part of a recently published open access paper (Jesus et al., 2023).

Methods. Nine (9) male and 8 female speakers from the same dialectal region in Portugal (*Dialetos Setentrionais* / North-western Dialects), aged 22 to 33 years (mean age of 26 years; standard deviation of 3 years) were recruited using convenience sampling. The participants were recorded in a quiet room, using a head-mounted Sennheiser Ear Set 1 condenser microphone, a sampling frequency of 48000 Hz and a bit depth of 16-bit per sample. Since no images of the glottal configurations were available at the time of data acquisition a Voice Specialist perceptually monitored and identified deviations from the targeted neutral whispering, described as normal adduction and medium loudness whisper (Konnai et al., 2017). Four sustained sibilants /s, z, ʃ, ʒ/ and 12 CVCV disyllabic real words with the same fricatives in initial, mid, and final word positions were used to estimate specific acoustic features of sibilants. These fricatives were also analysed in six sentences and a phonetically balanced text that are part of the speech materials used regularly in Portugal to evaluate voice quality. Multitaper power spectral density estimates based on 12 ms Hamming windows centred in the middle of the fricative were analysed using the slope of two regression lines (m1 – low frequencies; m2 – high frequencies), the broad peak frequency (F_{BP}) and broad peak level (L_{BP}). The fricative's median sound pressure level (SPL) over a 46 ms window, absolute and relative (to control for possible speech-rate effects) durations, were also calculated. Matlab 9.5.0.944444 (R2018b) and Praat 6.0.47 scripts were developed for signal processing and analysis; IBM SPSS Statistics 25, R version 4.3.1 running in RStudio Version 2023.06.1+524 and the beeswarm 0.4.0 package were used for statistical analysis, mixed-effects logistic regression modelling and data visualisation.

Results. No significant differences between m1 values of voiceless fricatives produced in the two speech modes (the exception being male's sustained /s/ and /s, ʃ/ in words), and significantly higher m1 values in female's whispered than voiced speech modes for (phonologically) voiced fricatives ($p < 0.010$; Student's t test; two-tailed p-values), except for the alveolar fricative /z/ in female's text and male's sentences. Place of articulation had a significant effect ($p < 0.010$; ANOVA with Bonferroni correction and Dunn's nonparametric comparison for post hoc testing after a Kruskal-Wallis tests; one-tailed p-values) on m1 values (the more posterior place of articulation had a steeper slope, i.e., higher m1 values), both in voiced and whispered speech modes, as shown in **Figure 1**. Results for m2 were not significantly different between the two speech modes, the only exceptions being: Sustained /s/, /s/ in words and text; female's sustained /z/; male's /ʃ/ in sentences; when /ʒ/ was produced in words. The values of F_{BP} for alveolar fricatives /s, z/ were significantly higher ($p = 0.000$; ANOVA with Bonferroni correction and Dunn's nonparametric comparison for post hoc testing after a Kruskal-Wallis tests; one-tailed p-values) than for postalveolar fricatives /ʃ, ʒ/, both for whispered and voiced speech

modes, in all four speech tasks and for both sexes, as shown in **Figure 1**. Voiceless fricatives were produced with a significantly higher L_{BP} value in voiced than in whispered speech mode, except for /ʃ/ produced by male speakers in sentences. Voiced fricative's L_{BP} results were not significantly different in the two speech modes, the only exception was /z/ produced in words by male speakers and /z, ʒ/ produced in words by female speakers. Whispered speech SPL was significantly lower than voiced speech, when the same fricative was compared in the two speech modes; this result held for both male and female speakers and the four speech tasks, except for /ʃ/ produced by male speakers in sentences. The absolute durations of same-place voiceless fricatives were only significantly different from voiced fricatives (/s/ versus /z/ and /ʃ/ versus /ʒ/) for the voiced speech mode. Nevertheless, the relative duration (shown in **Figure 1**) of same-place and speech mode voiceless fricatives was significantly higher ($p < 0.040$; Dunn's nonparametric comparison for post hoc testing after a Kruskal-Wallis test; one-tailed p-values) than voiced fricatives, except for female /ʃ/-/ʒ/ produced in text (both speech modes) and /s/-/z/ produced in whispered words, and male voiced /ʃ/-/ʒ/ in text.

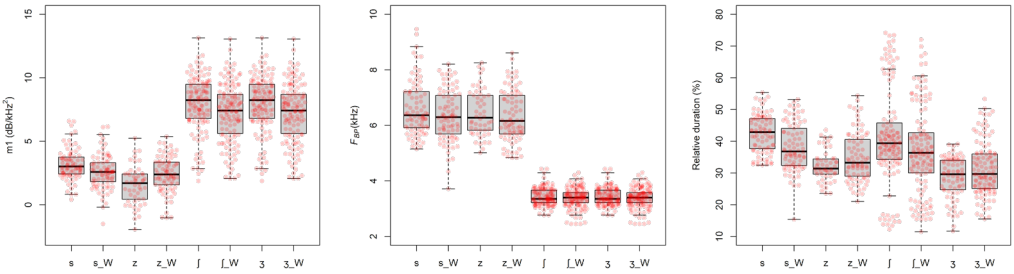


Figure 1: *Male voiced /s, z, ʃ, ʒ/ and whispered /s_W, z_W, ʃ_W, ʒ_W/ fricatives' low frequencies spectral slope (left), broad peak frequency (centre) and relative duration (right) in words.*

Discussion. m1 and F_{BP} are attributes associated with consonantal place of articulation and the relative duration carries sufficiently distinct information to disambiguate consonant voicing both in voiced and whispered speech. The fricatives' source strength (correlated with m1 values) was not significantly different between voiceless fricatives produced in the two speech modes and significantly different for voiced fricatives; place of articulation had a significant effect on source strength, both in voiced and whispered speech. The parameters (F_{BP} and L_{BP}) expected to correspond to the first front cavity resonance (fricative filter characteristics) revealed the same shifts in frequency (F_{BP}) with the place of articulation in whispered and voiced speech modes. Since L_{BP} is maximised for a higher source strength, our results constitute new cumulative evidence that voiceless fricatives are produced with a weaker source in whispered speech. The relative duration of same-place and speech mode voiceless fricatives was higher than voiced fricatives both in voiced and whispered speech, amounting to the only viable cue for voicing in whisper. This study, with data collected from different speech tasks, shows that changes during whispered speech production can be observed both in the laryngeal (source) and vocal tract (filter) configurations. Therefore, clinicians who use the whisper technique for voice rehabilitation, usually centred on the absence of vocal fold vibration, should also consider relevant changes in vocal tract configuration.

References

Heeren, W. F. L. (2015). Coding pitch differences in voiceless fricatives: Whispered relative to normal speech. *The Journal of the Acoustical Society of America*, *138*(6), 3427–3438. https://doi.org/10.1121/1.4936859

Jesus, L. M. T., Castilho, S., Ferreira, A., & Conceição Costa, M. (2023). Discriminative segmental cues to vowel height and consonantal place and voicing in whispered speech. *Journal of Phonetics*, *97*, 101223. https://doi.org/10.1016/J.WOCN.2023.101223

Jovičić, S. T., & Šarić, Z. (2008). Acoustic analysis of consonants in whispered speech. *Journal of Voice*, *22*(3), 263–274. https://doi.org/10.1016/j.jvoice.2006.08.012

Konnai, R., Scherer, R. C., Peplinski, A., & Ryan, K. (2017). Whisper and phonation: Aerodynamic comparisons across adduction and loudness. *Journal of Voice*, *31*(6), 773.e11-773.e20. https://doi.org/10.1016/j.jvoice.2017.02.016

Meynadier, Y. (2015). Aerodynamic tool for phonology of voicing. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*.

Murry, T., & Brown, W. S. (1976). Peak intraoral air pressures in whispered stop consonants. *Journal of Phonetics*, *4*(3), 183–187. https://doi.org/10.1016/S0095-4470(19)31242-2

Scherer, R. C., Sundberg, J., & Konnai, R. (2016). Whisper. In R. T. Sataloff & M. S. Benninger (Eds.), *Sataloff's Comprehensive Textbook of Otolaryngology: Head and Neck Surgery: Laryngology, Vol. 4.* (pp. 81–87). Jaypee Brothers Medical Publishers.

Sharifzadeh, H. R., McLoughlin, I., & Russell, M. J. (2012). A comprehensive vowel space for whispered speech. *Journal of Voice*, *26*(2), e49–e56. https://doi.org/10.1016/j.jvoice.2010.12.002

Slis, I. H., & Cohen, A. (1969). On the complex regulating the voiced-voiceless distinction I. *Language and Speech*, *12*(2), 80–102.

Sundberg, J., Scherer, R., Hess, M., & Müller, F. (2010). Whispering - A single-subject study of glottal configuration and aerodynamics. *Journal of Voice*, *24*(5), 574–584. https://doi.org/10.1016/j.jvoice.2009.01.001

Weismer, G., & Longstreth, D. (1980). Segmental gestures at the laryngeal level in whispered speech. *Journal of Speech, Language, and Hearing Research*, *23*(2), 383–392. https://doi.org/10.1044/jshr.2302.383

Zhang, C., & Hansen, J. (2007). Analysis and classification of speech mode: Whispered through shouted. *Proceedings of Interspeech 2007*, 2289–2292. https://doi.org/10.21437/Interspeech.2007-621

Zygis, M., Pape, D., Koenig, L. L., Jaskula, M., & Jesus, L. M. T. (2017). Segmental cues to intonation of statements and polar questions in whispered, semi-whispered and normal speech modes. *Journal of Phonetics*, *63*, 53–74. https://doi.org/10.1016/j.wocn.2017.04.001

C-G vs. C-V Timing Differences in Hong Kong Cantonese

Po-rong Chen[1], Feng-fan Hsieh[1], Yueh-chin Chang[1]

[1]National Tsing Hua University

perrychen1999@gmail.com, ffhsieh@mx.nthu.edu.tw, ycchang@mx.nthu.edu.tw

Introduction. This study addresses the ongoing debate surrounding the phonological status of labialized velar stops (*kw-* and *gw-*) in Hong Kong Cantonese (HKC). Traditionally considered co-articulated onsets (see, e.g., Bauer & Benedict 2011), their exact status as complex segments (/kw/) or sequences of separate phonemes (/k/+/w/) has yet to be conclusively established without instrumental investigation. Further complicating the issue is the existence of diphthongs like /ui/ in words like *bui* 'cup,' which are not rendered as *bwi* in *Jyutping* transliteration. To shed light on this controversy, we employ Shaw et al.'s (2021) heuristics for analyzing segmental compositions (see below). We hypothesize that the temporal organization of articulatory gestures will differ between complex segments and segment sequences, with C-G sequences (e.g., *kw-*) exhibiting tighter co-articulation and less temporal separation compared to C-V sequences (e.g., *ku-* or *bui*). Additionally, an examination is conducted on the stiffness (i.e., amplitude-normalized peak velocity; see, e.g., Roon et al. 2021) of glides and vowels as another potential differentiator (Burgdorf & Tilsen 2021). By focusing on the temporal organization and stiffness of specific articulatory gestures using Electromagnetic Articulography (EMA), this study offers an experimental perspective on this long-standing issue, thus contributing to a clearer understanding of syllable structure in HKC as well as the glide and vowel distinction in articulation.

Methods. Six Hong Kong Cantonese speakers (4 male) in their twenties participated in this study. The target syllables were embedded in disyllabic words (e.g., *bui1 dip6* 'cup (and) dish'). Three types of the target items are analyzed and reported: (1) labialized velar stops {*kwai*, *kwong, gwai*, and *gwaai*}, (2) syllables with diphthongs *ui* or *iu*: {*fui*, *bui*, *kui*, *piu*, *biu*, *tiu*, *diu*, *giu*, *siu*, and *ziu*}, and (3) syllables with monophthongs {*bun*, *gong*, *bing*, *ding*, and *ging*}. These syllables all bear Tone 1 (i.e., high level tone). Ten repetitions were collected for each target item. The target items were embedded in the carrier phrase: *gaa __ bei keoi* /kaa __ pei kʰøi/, meaning 'add __ for him/her.' The EMA data was processed using Mview (Tiede 2005) and custom MATLAB scripts. We used the vertical dimension of lower lip (LLz) trajectories to label the gestures of labial stops, the vertical dimension of tongue tip (TTz) to label the gestures of alveolars, and the vertical dimension of tongue dorsum (TDz) to label the gestures of velars. /i/ was analyzed using the vertical dimension of tongue blade (TBz) trajectories, while lip-rounding, including /u/, /w/ and /o/, was measured with the horizontal dimension of the upper lip (ULx). To quantify the stiffness, we employed the peak velocity divided by the amplitude (or, amplitude-normalized peak velocity), following Roon et al. (2021). Finally, the onset of *bei* 'for' from the carrier phrase served as the anchor for data normalization.

This study investigates the temporal organization of Cantonese C-G sequences, such as *kwai* and C-V sequences, such as *bui* or *bun*, focusing on their potential status as complex segments or segment sequences. Building upon Shaw et al. (2021)'s method, we analyze the relationship between two factors: the duration of the initial consonant plateau (G1) and the time difference between consonant and glide/vowel onsets (i.e., onset-to-onset: C-G/C-V lags). Their findings suggest that, in some languages like Russian, C-V/C-G lags remain unaffected by G1 duration for complex segments (hence palatalized consonants). In contrast, when onset-to-onset lags lengthen relative to G1 plateau duration, as seen in sequences such as *bj-* in American English, this indicates a mismatch in timing between *b* and *j*, thus suggesting a segment sequence rather than a single complex unit.

Results. Utilizing least-squares linear regression and correlation analysis, we examined the influence of G1 duration on onset-to-onset lags across different syllable types (C-G, C-V (diphthongs), and C-V (monophthongs)). As can be observed in Table 1, no significant correlation was found between G1 duration and G-G/C-V lag in C-G and C-V (diphthongs) syllables ($|r|<0.3$, indicating a weak correlation), including *gwai*, *bui*, and the like. This suggests that the onset consonant and the vocoids are timed in-phase, supporting their classification as complex segments since the onset-to-onset lags are not affected by the onset consonant plateau. In contrast, C-V (monophthongs) displayed varying degrees of gestural overlap, with syllables like *bun* and *gong* exhibiting moderate correlations ($0.3<|r|<0.7$) and others like *bing*, *ding*, and *ging* showing stronger correlations ($|r|>0.7$) between G1 duration and C-V (monophthongs) lag. This finding suggests that different vowel categories, such as front vowels versus back vowels, might influence patterns of gestural coordination within C-V sequences.

Analyzing amplitude-normalized peak velocity further revealed the influence of vowel categories on rapidity. Labio-velars like /u/ and /w/ (measured by ULx) generally exhibited higher peak velocities compared to palatals like /i/ and /j/ (TBz). Notably, while /u/ and /w/ showed no significant difference in stiffness, glide /j/ was significantly faster than vowel /i/ ($p<0.05$) in the present results.

Table 1: *Summary of regression coefficients (β), correlation coefficients (r) and stiffness*

C-G			C-V (diphthongs)			Amplitude-normalized peak velocity		
syllables	β	|*r*|	syllables	β	|*r*|	syllable types	average	*SD*
kwai	-0.2	0.14	*fui*	0.06	0.06	Glides (*w*)	27.77	14.15
gwai	0.09	0.06	*bui*	0.21	0.05	Diphthong (*u*)	30.33	16.86
gwaai	0.23	0.11	*kui*	0.23	0.16	Monophthong (*u*)	28.22	16.21
gwong	0.21	0.17	*piu*	-0.11	0.08	Diphthong (*i*)	22.87	11.95
C-V (monophthongs)			*biu*	0.37	0.28	Monophthong (*i*)	17.57	5.51
bun	0.35	0.33	*tiu*	0.24	0.24			
gong	0.34	0.33	*diu*	-0.03	0.04			
bing	1.06	0.71	*giu*	0.11	0.16			
ding	0.72	0.64	*siu*	0.38	0.26			
ging	1.02	0.89	*ziu* (*z:*/ts/)	0.07	0.29			

Discussion. Our results present a difference between C-G and C-V gestures in articulatory timing in Hong Kong Cantonese, although no substantial difference is found in rapidity (stiffness). To begin with, we compared C-G with C-V (monophthongs): the former constitutes complex segments (*gwai*), while the latter are comprised of segment sequences (*bun*). In other words, C-G gestures are timed in-phase, while C-V (monophthongs) gestures are timed anti-phase in HKC. Furthermore, C-V (diphthongs) gestures, such as *bui*, are also timed in-phase. In sum, CG (*gwai*) *v*s. CV (diphthongs) (*bui*) are complex segments, meaning that there is good evidence to suggest that these /w/'s and /u(i)/'s can be both analyzed as labialization. Interestingly, this stands in contrast to C-V (monophthongs) gestures, such as *bun*, which show anti-phase timing (see also Kramer et al. 2023 for similar results of C-V timing in Standard Chinese). Specifically, high correlation coefficients in C-V (monophthongs) gestures, particularly with high front vowels like {*bing*, *ding*, *ging*}, suggest a stronger degree of "mutual independence" compared to complex segments (e.g., C-V (diphthong) gestures like /biu/), which also exhibit in-phase timing. The stronger degree of "gestural separation" observed in C-V structures with high front vowels like {*bing*, *ding*, *ging*}, might be due to their greater coarticulatory resistance and aggressiveness, as reported by Recasens & Rodríguez (2016). Their study suggests that high front vowels exhibit stronger coarticulatory resistance effects on surrounding consonants, potentially hindering the complete gestural overlap of consonant and vowel.

Conclusion. This study investigated the temporal organization of Cantonese C-G and C-V sequences, focusing on their potential status as complex segments or segment sequences. Building upon Shaw et al.'s (2021) method, we analyzed the relationship between two factors: the duration of the initial consonant plateau (G1) and the time interval between consonant onset and vowel/glide onset (C-G/C-V lags). Our findings suggest that C-G sequences and C-V (diphthongs) exhibit in-phase timing, with no significant correlation between G1 duration and C-G/C-V lags, supporting their classification as complex segments. In contrast, C-V (monophthongs) displayed varying degrees of gestural separation, with high front vowels showing the strongest correlation between G1 duration and C-V lags. This suggests a more intricate gestural coordination within these sequences, possibly influenced by the coarticulatory resistance and aggressiveness of high front vowels reported in previous studies. Further research is needed to fully understand the factors shaping gestural patterns and segmental composition in C-V sequences, particularly those involving high front vowels.

In conclusion, the present results indicate that rising diphthongs conventionally transcribed as *wai, ui*, and *iu* are invariably complex segments (i.e., secondary articulation). Consequently, our finding challenges the intuition behind *Jyutping* transliteration, which treats these as different structures: either labialized velar stops (*gwai*) or diphthongs (*bui*). Needless to say, this mismatch between retrospective judgment and articulatory data warrants further investigation.

References

Bauer, R. S., & Benedict, P. K. (2011). *Modern Cantonese phonology* (Vol. 102). Walter de Gruyter.

Burgdorf, D. C., & Tilsen, S. (2021). Temporal differences between high vowels and glides are more robust than spatial differences. *Journal of Phonetics*, *88*, 101073.

Kramer, B. M., Stern, M. C., Wang, Y., Liu, Y., & Shaw, J. A. (2023) Synchrony and Stability of Articulatory Landmarks in English and Mandarin CV Sequences. In: Radek Skarnitzl & Jan Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1022–1026). Guarant International.

Recasens, D., & Rodríguez, C. (2016). A study on coarticulatory resistance and aggressiveness for front lingual consonants and vowels using ultrasound. *Journal of Phonetics*, *59*, 58-75.

Roon, K. D., P. Hoole, C. Zeroual, S. H. Du, and A. I. Gafos. 2021. Stiffness and articulatory overlap in Moroccan Arabic consonant clusters. *Laboratory Phonology* 12. DOI: 810.5334/labphon.272.

Shaw, J. A., Oh, S., Durvasula, K., & Kochetov, A. (2021). Articulatory coordination distinguishes complex segments from segment sequences. *Phonology*, *38*(3), 437-477.

Tiede, M. (2005). MVIEW: software for visualization and analysis of concurrently recorded movement data. New Haven, CT: Haskins Laboratories

Poster 4

11:00 am – 01:00 pm

Paper	Title	Authors
59	Features Used to Discriminate Vowel Height in Voiced and Whispered Speech	Luis Jesus (University of Aveiro)*; Sara Castilho (Hospital Arcebispo João Crisóstomo, Cantanhede); Aníbal JS Ferreira (University of Porto - Faculty of Engineering); Maria Conceição Costa (University of Aveiro)
156	Examining the Link between the Perception and Production of Phonetic Convergence of Laughter in Interaction	Marin Schröer (Universität Bielefeld)*; Bogdan Ludusan (Bielefeld University)
249	Phonetic convergence in French conversational interaction	Cyrielle Mack (Aix-Marseille University); Pauline Tudose (Aix-Marseille University); Noel Nguyen (Aix-Marseille University)*
126	Testing the generality of L1 phonetic precision as a predictor of L2 VOT acquisition	Marie K Huffman (Stony Brook University)* ; Katharina Schuhmann (Carl von Ossietzky Universität Oldenburg)
245	Malleability of speech sound representations: bite blocks and after effects	Xinyu Zhang (Radboud University)*; Rob Schoonen (Radboud University); Esther Janse (Radboud University Nijmegen)
205	Explorations into speaker consistency in speech breathing and anticipation of upcoming phonetic content	Laura L. Koenig (Adelphi University, Haskins Labs)*; Susanne Fuchs (zas)
26 (Remote)	COMPARISON OF ACOUSTIC AND PHYSIOLOGICAL MEASURES OF COARTICULATION	Irfana Madathodiyil (All India Institute of Speech and Hearing, Manasagangothri)* and Fathima Nuha (All India Institute of Speech and Hearing, Manasagangothri)
134	Simulation-based Bayesian inference of state feedback control model parameters to fit f0 perturbation responses in laryngeal dystonia	Jessica L Gaines (UC Berkeley - UCSF Graduate Program in Bioengineering)*; Kwang S Kim (Purdue University); Ben Parrell (University of Wisconsin-Madison); Vikram Ramanarayanan (University of California, San Francisco & Modality.AI); Alvince L Pongos (Berkeley); Srikantan Nagarajan (UCSF); John F Houde (University of California San Francisco)
61	Analysis of tongue movement (quasi)-â€œsteady-statesâ€ using General Tau theory	Alice Turk (The University of Edinburgh)*; Benjamin Elie (The University of Edinburgh); Cedric Macmartin (The University of Edinburgh); David N. Lee (The University of Edinburgh)
164	Brain changes associated with stuttering therapy and spontaneous recovery	Nicole E Neef (University Medical Center Goettingen)*; Soo-Eun Chang (University of Michigan)
22	Segmental durations and the vowel length contrast in fast speech in Hungarian	Andrea Deme (ELTE Eötvös Loránd University)*; Kornélia Juhász (ELTE Eötvös Loránd University & HUN-REN Hungarian Research Centre for Linguistics); Zsuzsa Szánthó (ELTE Eötvös Loránd University); Szabina Zsoldos (ELTE Eötvös Loránd University); Reinhold Greisbach (University of Cologne)
100	Does the ultrasound probe affect articulatory gesture in children? An acoustic study.	Laura Machart (Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France *Institute of Engineering)*; anne vilain (UGA/GIPSA-lab); Hélène Loevenbruck (Univ. Grenoble Alpes, CNRS, LPN, 38000 Grenoble, France); Lucie Ménard (Université du Québec à Montréal)
2	SPRAAKLAB: mobile laboratory to collect high-quality speech data	Martijn Wieling (University of Groningen)*; Teja Rebernik (University of Groningen); Jidde Jacobi (University of Groningen); Thomas B Tienkamp (University of Groningen); Frank Tsiwah (University of Groningen); Defne Abur (University of Groningen)
157	Introducing ADA: A Tool for Articulatory Data Analysis	Philipp Buech (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne Nouvelle)*; Anne Hermes (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS & Sorbonne Nouvelle, Paris)
221	On The Utility of a Single-Breath Counting Task for the Remote Digital Assessment of Respiratory Function in ALS	Michael Neumann (Modality.AI, Inc.)*; Hardik Kothare (Modality.AI); Vikram Ramanarayanan (University of California, San Francisco & Modality.AI)
48	Auditory-Motor Adaptation of Vowels Across Adulthood	Katharina M. Polsterer (University of Groningen)*; Thomas B Tienkamp (University of Groningen); Teja Rebernik (University of Groningen); Hedwig Sekeres (University of Groningen); Valentine Lucquiaud (University of Groningen); Martijn Wieling (University of Groningen); Defne Abur (University of Groningen)
148	Onset-cluster production in Mandarin according to sonority profile	Xuejing Chen (Laboratoire de Phonétique et Phonologie (CNRS & U. Sorbonne Nouvelle))*; Pierre A Hallé (CNRS); Rachid Ridouane (LPP (CNRS & Sorbonne Nouvelle))
110	praatpicture: A library for making flexible Praat Picture-style figures in R	Rasmus Puggaard-Rode (Institute for Phonetics and Speech Processing, LMU Munich)*
97	Tongue root movement in Hungarian intervocalic alveolar obstruents	Tekla Etelka Gráczi (Hungarian Research Centre for Linguistics)*; Kornélia Juhász (HUN-REN Hungarian Research Centre for Linguistics); Péter Csényi (HUN-REN Hungarian Research Centre for Linguistics); Tamás G Csapó (Budapest University of Technology and Economics); Andrea Deme (ELTE Eötvös Loránd University); Alexandra Markó (MTA-ELTE Lingual Articulation Research Group)
32 (Remote)	Acoustic correlates of the nasal vs. plosive quantity contrast in Hungarian	Tilda Neuberger (HUN-REN Hungarian Research Centre for Linguistics)*
178	Speech and language abilities associated with regional corpus callosum development in children who stutter	Fiona Hübler (University of Michigan)*; Emily Garnett (University of Michigan); Yanni Liu (University of Michigan); Ho Ming Chow (University of Delaware); Soo-Eun Chang (University of Michigan)
159	Phoneme monitoring and articulatory suppression in French-speaking adults	Claire Boilley (Université Grenoble Alpes)*; anne vilain (UGA/GIPSA-lab); Patricia Pires (Université Grenoble Alpes)
174	Dimensions of structure and variability in the human vocal tract	Katherine Vaughan-Williams (Lancaster University); Steven Moran (University of Neuchâtel / University of Miami); Sam Kirkham (Lancaster University)*
90	Production and perception of tonal coarticulation: Evidence from computational simulation of communication	Po-Hsuan Huang (National Taiwan University)*
202	Inter-gestural coupling of onset and vowel gestures in adults who stutter in different rhythmic conditions	Mona Franke (Institute of Phonetics and Speech Processing)*; Simone Falk (University of Montreal); Philip A Hoole (Institute of Phonetics, Munich University)

195	How Do Speakers Respond to Altered Formant Feedback Simulating a Change in Speaker Gender?	Erin Daly (New York University); Douglas Shiller (Université de Montréal); Vesna Novak (University of Cincinnati); Tara McAllister (New York University)*
166	Control and Recovery of Vocal Tract Gestures Using Stochastic Models of Action and Perception: The Martingale Dynamics of Human Speech	Gordon Ramsay (Emory University)*
146	Spatiotemporal features of bilabial geminate and singleton consonants in Italian	Francesco Burroni (LMU Munich)*; Sireemas Maspong (LMU Munich); Nicole Benker (LMU Munich); Philip A Hoole (Institute of Phonetics, Munich University); James Kirby (Institute of Phonetics and Speech Processing)
7	Are "frequency" effects cumulative from word to syllable?	Ivan Yuen (Saarland University)*; Bistra Andreeva (Saarland University); Omnia Ibrahim (Saarland University); Bernd Möbius (Saarland University)
25	Music in the treatment of childhood motor speech disorders: Using music to cue gestural timing	Mirjam v Tellingen (Rehabilitation Center 'Revalidatie Friesland')*; Jozet Nurkowara (Rehabilitation Center 'Revalidatie Friesland'); Anne Marie van de Zande (Rehabilitation Center 'Rijndam Revalidatie'); Maya Verband (Department of Communication Sciences and Disorders, University of Iowa); Ben A. Maassen (Center for Language and Cognition Groningen, University Medical Center Groningen); Roel Jonkers (Center for Language and Cognition Groningen, University of Groningen)
115	Merging verb forms with "ich" to enchaînement consonantique in German	Jürgen Trouvain (Saarland University)*; Christine Mooshammer (Humboldt-Universität zu Berlin); Malte Belz (Humboldt-Universität zu Berlin); Robert Lange (Humboldt-Universität zu Berlin)
[illegible]	[illegible]	[illegible]
33 (Remote)	Use of Natural Anchors for Improving Rater Reliability in Dysarthria Assessment: An Exploratory Study	Mili Kuruvilla-Dugdale (University of Iowa)*; Thushara Manasinghe (University of Iowa); Dorothy Cresta (University of Iowa); Keila L Stigerwalt (University at Buffalo); Amy Arel (University of Iowa)
127	Muscle synergies in the production of stop consonants with increasing intensity	Maeva GARNIER (GIPSA-lab)*; Marion Léger (GIPSA-lab); Julien Frère (GIPSA-lab)
214	Effects of Aging on /s/ in Spontaneous Speech Using Neural Networks and Phonetic Measures	Oriane Dufour (Laboratoire de phonétique et phonologie)*; Anne Hermes (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS & Sorbonne Nouvelle, Paris); Cédric Gendrot (LPP)
196	Acoustic and kinematic correlates of adaptive responses to consistent formant perturbations in young healthy speakers	Teja Rebernik (University of Groningen); Tomas Lentz (Tilburg University); Maya Verband (Department of Communication Sciences and Disorders, University of Iowa)*
208	LSTM analysis of time scales of breakdown in ALS	Jessica A Campbell (University of Southern California)*; Louis Goldstein (University of Southern California); Khalil Iskarous (University of Southern California)
215	Exploring F0 Entrainment in Bi-directional Speech Imitation: A Cross-Linguistic Analysis	Zhang Yuan (Istituto Italiano di Tecnologia)*
20	Insights into phonemes' articulation time	Montserrat Sobaranes (UNAM)*; Carlos A. Pérez-Ramírez (UAQ); M. Florencia Assaneo (UNAM)
18	Exploration and classification of vocal fry, period doubling, and modal voice using acoustic and EGG measures	Yaqian Huang (Acoustics Research Institute, Austrian Academy of Sciences)*
91	Perception of a four-way stop laryngeal contrast in Eastern Oromo	Maida Percival (University of Toronto)*
27 (Remote)	ARTICULATORY DYNAMICS IN A TONAL LANGUAGE	L Tamil Dangulwara (All India Institute of Speech and Hearing, Manasagangothri) and Irfana Madathodiyil (All India Institute of Speech and Hearing, Manasagangothri)*
79	Lingual articulation of syllabic and non-syllabic /r/	Kateřina Bujalová (Charles University)*; Tanja Kocjančič Antolík (Charles University)*
154	Positional Effect of Main Stress in Italian: an Articulatory and Acoustic Study	Baosi Shao (Département d'études cognitives, Ecole Normale Supérieure - Université PSL)*; Philipp Buech (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Sorbonne Nouvelle); Anne Hermes (Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS & Sorbonne Nouvelle, Paris); Maria Giavazzi (Département d'études cognitives, Ecole Normale Supérieure - Université PSL)
16	Articulatory planning of spoken utterances based on Optimal Control Theory	Benjamin Elie (University of Edinburgh)*, Juraj Simko (University of Helsinki), Alice Turk (The University of Edinburgh)
19	Chewing Efficiency and Oral developmental functions in Children with Oral- and Speech Motor Disorders Compared to Peers	Helena Björklin (Karolinska Institutet, Department of Clinical Sciences, Danderyd Hospital, Stockholm)*; Maya Verband (Department of Communication Sciences and Disorders, University of Iowa); Jenny Trang (Karolinska Institutet, Department of Dental Medicine, Division of orthodontics and pediatric dentistry, Stockholm); Fredrik Johansson (Medical library, Danderyd Hospital, Stockholm); Georgios Tsilingaridis (Karolinska Institutet, Department of Dental Medicine, Division of orthodontics and pediatric dentistry, Stockholm); Royne Thurman (Karolinska Institutet, Department of Clinical Sciences, Danderyd Hospital, Stockholm)
152	Timing of acceleration peaks and acceleration changes	Malin Svensson Lundmark (Lund University)*

Features Used to Discriminate Vowel Height in Voiced and Whispered Speech

Luis M. T. Jesus [1], Sara Castilho[2], Aníbal J. S. Ferreira[3], Maria Conceição Costa[4]

[1]School of Health Sciences (ESSUA), Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Intelligent Systems Associate Laboratory (LASI), University of Aveiro, Portugal
[2]Hospital Arcebispo João Crisóstomo, Cantanhede, Portugal
[3]Department of Electrical and Computer Engineering, University of Porto, Portugal
[4]Department of Mathematics (DMat) and Centre of Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal

lmtj@ua.pt, sara.castilho@ua.pt, ajf@fe.up.pt, lopescosta@ua.pt

Introduction. Acoustic studies of vowels have shown that F_1 and F_2 frequencies are higher in whispered speech than in voiced speech (Maurer, 2016; Swerdlin et al., 2010). Matsuda and Kasuya (1999) found that models incorporating weak acoustic coupling between the subglottal system and a constriction between the false vocal folds, can simulate this raising of the frequency of lower formants observed in whispered speech. Furthermore, Sharifzadeh et al. (2012) found that whispered /ə/ and /ʌ/ formant frequency shifts from voiced reference values were more pronounced than for other vowels. In whispered vowels there was also more convergence of adjacent vowels, for example, /i/ and /ɪ/ F_1 and F_2 frequency values were more similar in whispered speech than in voiced speech (Sharifzadeh et al., 2012). Duration and fundamental frequency (f_o) are also used as complementary (to formant frequencies) features to discriminate vowels (Heeren, 2015). In whispered speech, formant frequencies, intensity and duration carry prosodic information. Intrinsic f_o has been shown (Jacewicz & Fox, 2015) to be positively correlated to vowel height, a phenomenon that plays out across more than 30 languages (Whalen & Levitt, 1995). Open vowels have been shown to be longer than close vowels, and height-related vowel duration differences are used in different languages as a secondary feature to enhance contrast (Cho, 2015). Vowels' intrinsic duration is also conditioned by physiological factors (Holt et al., 2015): Vowels that are produced with a low jaw are longer than those produced with high jaw position. In this paper we compare the characteristics of voiced and whispered vowels in different speech tasks, produced by speakers from the same dialectal region and age group. Our aim was to identify which height cues are used consistently across the two speech modes (voiced and whispered). This work has been previously published as part of an open access paper (Jesus et al., 2023).

Methods. Seventeen (17) participants (9 male speakers and 8 female speakers; 22 to 33 years of age) were recruited using convenience sampling in the districts of Aveiro and Coimbra in Portugal. Participants were seated in a quiet room and recorded using a head mounted condenser microphone. Acoustic data was sampled at 48000 Hz with 16 bits per sample. A similar screening and training procedure to that previously used (Konnai et al., 2017) to ensure participants can discriminate and produce voiced and whispered speech was adopted in this study. Materials included four sustained oral vowels, 12 CVCV disyllabic real words, six sentences used by clinicians to evaluate voice quality and a phonetically balanced text. We only analysed the four oral vowels /i, a, ɔ, u/ that define the corners of the EP vowel space (Escudero et al., 2009). The parameters used to analyse the vowels were: f_o; spectral slope; sound pressure level (SPL); F_1, F_2 and F_3 frequencies. We also extracted absolute durations as in previous studies (Escudero et al., 2009), and calculated the following relative durations to control for possible speech-rate effects: Phone to word-length ratio of the word task; phone to sentence-length ratio in the sentence reading task; phone to text-length ratio (including pauses) in the phonetically balanced text reading task. Matlab 9.5.0.944444 (R2018b) and Praat 6.0.47 scripts were developed for signal processing and analysis; IBM SPSS Statistics 25, R version 4.3.1 running in RStudio Version 2023.06.1+524 and the beeswarm 0.4.0 package were used for statistical analysis, mixed-effects logistic regression modelling and data visualisation.

Results. A significant positive correlation between voiced and whispered F_1 frequencies (shown in **Figure 1**) of female (Spearman's correlation coefficient = 0.924, p = 0.000) and male (Spearman's correlation coefficient = 0.947, p = 0.000) speakers was observed. The same positive correlation was found to be significant between voiced and whispered F_2 frequencies of female (Spearman's correlation coefficient = 0.994, p = 0.000) and male (Spearman's correlation coefficient = 0.979, p = 0.000) speakers. A significant positive correlation was also found between voiced and whispered F_3 frequencies, both for female (Spearman's correlation coefficient = 0.921, p = 0.000) and male (Spearman's correlation coefficient = 0.691, p = 0.003) speakers. The vowel space area calculated using a polygon with vertices at the mean value for each vowel (McCloy, 2016), revealed a compression in whispered speech, when compared to an equivalent voiced speech task, both for female and male speakers. A clear downward shift (relative to voiced speech) of vowel spaces in whispered speech could be observed for all speech tasks. Female and male speakers' spectral slope values of all vowels increased significantly (Student's t and Mann-Whitney U tests) for whispered speech (relative to voiced speech), and spectral slope findings were consistent across tasks. The SPL of all of female's and male's whispered vowels was

significantly lower than in voiced exemplars, with a mean downward shift between 19 and 25 dB, that was very stable across speech tasks. A significant positive correlation was found for f_o and $F_{1(whispered)}$ - $F_{1(voiced)}$ of female speakers (Pearson's correlation coefficient = 0.660, p = 0.005; two-tailed p-value). Absolute durations of female and male voiced and whispered speech were used to differentiate close /i, u/ from open/open-mid /a, ɔ/ vowels, the only exception being the values of male /i/ when compared to /ɔ/. A Kruskal-Wallis test provided evidence of a difference (p = 0.000) between the mean ranks of at least one pair of groups of all the different possible multi-comparisons. Dunn's pairwise tests of female and male, voiced and whispered speech were carried out for the four pairs (/i/-/a/; /i/-/ɔ/; /u/-/ɔ/; /u/-/a/), showing significantly different durations between close and open/open-mid vowels, except for male voiced /i/-/ɔ/ (p = 0.152) and /u/-/ɔ/ (p = 0.195) pairs. The relative durations shown in **Figure 1** correspond to the phone to text-length ratio of the text task. Both female and male relative durations (voiced and whispered speech) unveiled a new pattern that had only just surfaced when looking at the absolute values: Close /i, u/ vowels were significantly shorter than open-mid vowels /a, ɔ/. A Kruskal-Wallis test provided evidence of a significant difference (p = 0.000; two-tailed p-value) between the mean ranks of at least one pair of groups. Dunn's pairwise tests were carried out for the four pairs (/i/-/a/; /i/-/ɔ/; /u/-/ɔ/; /u/-/a/). There was evidence, that intrinsic vowel durations were at play here, even when the speakers whispered the vowels.

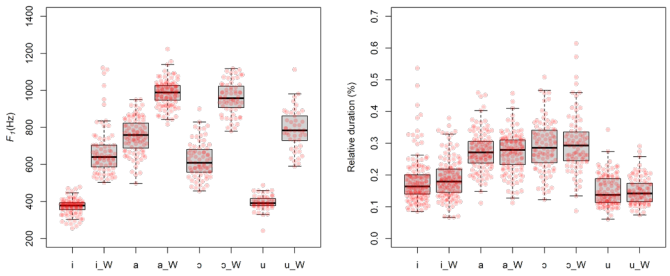


Figure 1: *Female voiced /i, a, ɔ, u/ and whispered /i_W, a_W, ɔ_W, u_W/ vowels' F_1 frequencies (left) and relative durations (right) in a phonetically balanced text.*

Discussion. Clear evidence has been found supporting that vowels are produced with significantly different F_1, F_2, spectral slope and SPL in voiced and whispered speech. A positive correlation between f_o values and F_1 shifts, relative to same-sex reference voiced F_1, in whispered speech was only found when analysing all the female tasks together. Close /i, u/ vowels durations were significantly shorter than close/open-mid vowels /a, ɔ/ both in voiced and whispered speech. The back cavity is likely to be shorter in whispered speech because the close-front unrounded vowels' Helmholtz resonance and the open-front unrounded back cavity resonance frequency were both significantly higher in whispered speech than in voiced speech mode. This may result from raising of the larynx and narrowing of the vocal tract around the ventricular folds for whispered speech production. F_1 frequency and relative duration were consistently used as height cues across the two speech modes (voiced and whispered). The evidence presented in this paper can be used to inform signal processing-based algorithms aiming at restoring voicing in whispered speech signals, and to inform rehabilitation strategies.

References

Cho, T. (2015). Language effects on timing at the segmental and suprasegmental levels. In M. A. Redford (Ed.), *The Handbook of Speech Production* (pp. 505–529). Wiley.

Escudero, P., Boersma, P., Rauber, A. S., & Bion, R. A. H. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America*, *126*(3), 1379–1393.

Heeren, W. F. L. (2015). Coding pitch differences in voiceless fricatives: Whispered relative to normal speech. *The Journal of the Acoustical Society of America*, *138*(6), 3427–3438. https://doi.org/10.1121/1.4936859

Holt, Y. F., Jacewicz, E., & Fox, R. A. (2015). Variation in vowel duration among southern African American english speakers. *American Journal of Speech-Language Pathology*, *24*(3), 460–469.

Jacewicz, E., & Fox, R. A. (2015). Intrinsic fundamental frequency of vowels is moderated by regional dialect. *The Journal of the Acoustical Society of America*, *138*(4), EL405–EL410. https://doi.org/10.1121/1.4934178

Jesus, L. M. T., Castilho, S., Ferreira, A., & Conceição Costa, M. (2023). Discriminative segmental cues to vowel height and consonantal place and voicing in whispered speech. *Journal of Phonetics*, *97*, 101223. https://doi.org/10.1016/J.WOCN.2023.101223

Konnai, R., Scherer, R. C., Peplinski, A., & Ryan, K. (2017). Whisper and phonation: Aerodynamic comparisons across adduction and loudness. *Journal of Voice*, *31*(6), 773.e11-773.e20. https://doi.org/10.1016/j.jvoice.2017.02.016

Matsuda, M., & Kasuya, H. (1999). Acoustic nature of the whisper. *Proceedings of Eurospeech 99*, 133–136.

Maurer, D. (2016). *Acoustics of the Vowel: Preliminaries*. Peter Lang.

McCloy, D. (2016). *Normalizing and Plotting Vowels with phonR 1.0.7*. University of Washington, USA. https://drammock.github.io/phonR/

Sharifzadeh, H. R., McLoughlin, I., & Russell, M. J. (2012). A comprehensive vowel space for whispered speech. *Journal of Voice*, *26*(2), e49–e56. https://doi.org/10.1016/j.jvoice.2010.12.002

Swerdlin, Y., Smith, J., & Wolfe, J. (2010). The effect of whisper and creak vocal mechanisms on vocal tract resonances. *The Journal of the Acoustical Society of America*, *127*(4), 2590–2598. https://doi.org/10.1121/1.3316288

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics*, *23*(3), 349–366. https://doi.org/10.1016/S0095-4470(95)80165-0

Examining the Link between the Perception and Production of Phonetic Convergence of Laughter in Interaction

Marin Schröer, Bogdan Ludusan

Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University, Germany
{marin.schroeer, bogdan.ludusan}@uni-bielefeld.de

Introduction. Phonetic convergence, the increase in similarity in the acoustic realizations of speakers, has been a widely researched topic in recent years (Pardo 2013). Evidence for convergence of most phonetic aspects of speech has been found, spanning different characteristics, such as vowel quality (Pardo et al. 2017), voice quality or speech rate (Levitan and Hirschberg 2011). While most of these studies focused on the verbal message (lexical items), there is evidence showing that convergence may occur also for non-verbal phenomena, such as laughter (Truong and Trouvain 2012).
Previous studies have employed different approaches to test phonetic convergence, which may be grouped into two classes: either focusing on production, by measuring and comparing distances between acoustic cues, or on perception, by asking raters to judge the similarity between certain audio stimuli (see Pardo 2013 for a review). Recently, there has been an increased interest to connect these two types of approaches, with studies incorporating both perceptual data and acoustic measurements (Pardo et al. 2017; Lewandowski and Nygaard 2018; M. Wagner et al. 2021). Having previously shown that laughter vowels undergo convergence in terms of their quality, measured based on distances between the values of the first two formants (Ludusan, Schröer, and P. Wagner 2022), we would also like to test this result perceptually and to analyze the link between production and perception, in the case of laughter in interaction.

Methods. We annotated the vowels of laughter instances from two interlocutors of a dialogue from the DUEL corpus (Hough et al. 2016) and then we manually extracted vowels from the beginning (first third) and end (last third) of the conversation, to be used as stimuli in the perception experiment. All but one were longer than 100 ms (mean 184 ms).
23 participants were presented the stimuli in a modified ABX paradigm, where the X stimulus was presented after each A and B. They were asked to rate which of the resulting pairs (AX or BX) was more similar. The stimuli had A taken from either the first or last third from one speaker, B taken from the opposite third of the same speaker and X from the first third of the opposite speaker. This was done in order to assess whether the speaker (S1 or S2) had converged to their interlocutor's baseline or not. In total, participants were presented 80 tokens of the structure described above.
We also extracted the mean value of different acoustic cues from the stimuli used in the perception experiment: the first and second formants (F1 and F2, respectively, to determine vowel quality), the fundamental frequency (*f0*), the root-mean-square energy of the signal (*en*), the duration of the vowel (*dur*) and the cepstral peak prominence (*cpp*, a measure of voice quality, with lower values of this measure indicating a more breathy phonation). We then computed the distances between vowels in each pair of stimuli, the Euclidean distance for F1 and F2 ($d_{f1f2}(A,B) = \sqrt{(F1_A - F1_B)^2 + (F2_A - F2_B)^2}$) and the absolute distance for all other cues (e.g., $d_{f0}(A,B) = |f0_A - f0_B|$).
We determined whether raters perceived convergence by computing the proportion of times they chose as being more similar the stimulus pair consisting of a vowel of a speaker from the last third and a vowel from the first third, produced by their interlocutor, and testing if this was different from chance level (50%). A significantly higher value would represent convergence, while a lower value would indicate divergence. The analysis was conducted on a per-speaker basis.
Acoustic-phonetic convergence was tested by means of Wilcoxon signed rank tests, determining whether the distances between stimuli from different interlocutors, from the first third, were different from those between a stimulus from the last third and one from the first third (again from different speakers), for each speaker separately. Both for the perceptual and for the acoustic analyses we chose a non-parametric test, as most distributions were non-normal.
Finally, in order to examine the link between perception and production, we fitted a generalized mixed effects model with the pooled data from both speakers. The dependent variable was the answer given by the raters: 1, for having chosen as more similar the pair containing stimuli from different thirds (the convergence case, and 0, otherwise). All cues and their interactions were fixed factors in the model, while we considered the rater as random intercept. The maximal model

was first built, and then reduced by removing a factor, one step at a time, as long as it reduced the Akaike Information Criterion value. The acoustic cues were normalized by subtracting their mean and dividing by two standard deviations.

Results. The participants in the perceptual experiment rated the stimuli from speaker S1 as showing convergence (58.9% convergence answer, $p = 2.0e^{-4}$), while those for speaker S2 as showing divergence (30%, $p = 2.8e^{-5}$).
We then examined whether there are differences between the the acoustic distances of the two pairs of stimuli. For speaker S1, there were significant differences for *dur* ($p = 0.003$) and for the *f1f2* distance ($p = 0.012$), both indicating convergence. For speaker S2, the acoustic analysis showed a more complex picture, with the values for *dur* ($p = 0.033$), *en* ($p = 6.8e^{-4}$) and *cpp* ($p = 1.9e^{-4}$) showing convergence, while those for *f0* ($p = 1.0e^{-7}$) showing divergence. The *f1f2* distance showed a trend towards convergence, although it was not significant ($p = 0.087$).
The model fitted to study the relation between raters' perception and stimuli acoustic distances reveled a significant main effect for *dur* ($\beta = 0.688, p = 1.2e^{-7}$), *f0* ($\beta = 1.516, p < 2.2e^{-16}$), *cpp* ($\beta = 0.424, p = 1.6e^{-3}$) and *f1f2* ($\beta = 0.933, p = 4.8e^{-9}$). There was no significant main effect for *en* ($\beta = 0.255, p = 0.060$). One two-way interaction (*en:f1f2*), four three-way interactions (*dur:en:cpp*, *dur:en:f1f2*, *en:f0:f1f2* and *f0:cpp:f1f2*), all but one of the four-way interactions (*dur:f0:cpp:f1f2*), and the five-way interaction were found to be significant.

Discussion. Investigating the role of several acoustic cues on perception, our study revealed that, for each of the examined cues, having a higher distance at the beginning of the conversation, compared to at its end, increased the odds of the raters to perceive convergence. While these findings may suggest a straightforward link between production and perception, in the case of phonetic convergence of laughter, a high number of interactions were found to be significant and many of them had a negative estimate. Thus, these results point towards a more complex picture, in which also the interactions between several acoustic cues need to be taken into account. Moreover, based on the fitted model we are able to draw conclusions on the importance of each acoustic cue for the perception of convergence, with the fundamental frequency of the voice playing the most important role, followed by vowel quality (as given by the Euclidean distance between the first two formant values), duration, voice quality (breathiness – as given by *cpp*) and finally, speech intensity.
The different importance ranking of the examined acoustic cues may explain the more complex case we encountered for speaker S2, where the duration, the energy of the signal, and the cepstral peak prominence measures indicated convergence, while f0 indicated divergence. Considering that the cue that played the most important role in perception (f0) showed divergence, it may not be surprising that the raters judged that speaker as diverging. These results do not fully align with those of a previous acoustic study of phonetic convergence (Ludusan, Schröer, and P. Wagner 2022), in which both speakers of this pair showed convergence. The difference is most likely due to the small subset of stimuli that were included in the current study, which might not be representative of the full set considered in the previous work (which analyzed a set one order of magnitude larger). In particular, some of the stimuli employed here had high f0 values, diverging from other stimuli, counteracting the convergence (or convergence trends) seen with respect to other measures.
These findings align though with those of previous studies examining production and perception aspects of convergence, for instance that f0, duration and vowel spectra are important predictors for convergence (Lewandowski and Nygaard 2018) and that the complex interaction of multiple acoustic cues is needed to fully account for convergence (Pardo et al. 2017). Our results further extend those of these studies, by showing that they hold also for non-verbal phenomena.

References.

Hough, Julian, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg (2016). "DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter". In: *Proc. of LREC*, pp. 1784–1788.

Levitan, Rivka and Julia Hirschberg (2011). "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions". In: *Proc. Interspeech 2011*, pp. 3081–3084. DOI: 10.21437/Interspeech.2011-771.

Lewandowski, Eva and Lynne Nygaard (Aug. 2018). "Vocal alignment to native and non-native speakers of English". In: *The Journal of the Acoustical Society of America* 144, pp. 620–633. DOI: 10.1121/1.5038567.

Ludusan, Bogdan, Marin Schröer, and Petra Wagner (2022). "Investigating phonetic convergence of laughter in conversation". In: *Proc. of INTERSPEECH*, pp. 1332–1336. DOI: 10.21437/Interspeech.2022-10332.

Pardo, Jennifer (2013). "Measuring phonetic convergence in speech production". In: *Frontiers in Psychology* 4, p. 559.

Pardo, Jennifer, Adelya Urmanche, Sherilyn Wilman, and Jaclyn Wiener (2017). "Phonetic convergence across multiple measures and model talkers". In: *Attention, Perception, & Psychophysics* 79, pp. 637–659. DOI: 10.3758/s13414-016-1226-0.

Truong, Khiet P. and Jürgen Trouvain (2012). "On the acoustics of overlapping laughter in conversational speech". In: *Proc. Interspeech 2012*, pp. 851–854. DOI: 10.21437/Interspeech.2012-192.

Wagner, Mónica, Mirjam Broersma, James McQueen, Sara Dhaene, and Kristin Lemhöfer (2021). "Phonetic convergence to non-native speech: Acoustic and perceptual evidence". In: *Journal of Phonetics* 88, p. 101076.

Phonetic convergence in French conversational interaction

Cyrielle Mack[1], Pauline Tudose[1], Noël Nguyen[1]

[1]Aix-Marseille Université, CNRS, LPL, Aix-en-Provence

cyrielle.mack@etu.univ-amu.fr, pauline.tudose@etu.univ-amu.fr, noel.nguyen-trong@univ-amu.fr

Introduction. Conversation is a social phenomenon with implications extending beyond the mere perception of an interlocutor's speech production, encompassing a complex interplay of cognitive, linguistic, and social processes. Successfully navigating these complexities requires constant monitoring, adaptation, and responsiveness to the dynamic nature of conversation. Convergence, and more specifically phonetic convergence, has been identified as one way to promote success in these endeavours. Studies in the last decades have shown that both speech shadowing and conversational interaction drive speech adjustment among speakers, fostering increased similarities in phonetic attributes (Fowler et *al.*, 2003; Goldinger, 1998; Namy et *al*., 2002; Pardo, 2006). It aligns with the principles of the Communication Accommodation Theory (Giles et *al*., 1991), a sociolinguistic framework positing that individuals strategically adjust their communication styles, either converging with or diverging from the speech patterns of their interlocutors. Trying to uncover the mechanisms driving phonetic convergence, as a key concept of this theory, will then help in better characterising the dynamic nature of language adaptation during conversation.

This research effort managed to weigh in on the still prominent debate opposing a one-way, automatic and direct model linking speech perception and production (see Pickering & Garrod, 2004, 2006) to an account more accommodating of the observed influence of both linguistic, including phonetic, and social factors.

Pardo (2006) laid the groundwork for investigating these questions by assessing how speakers enhance phonetic similarities during conversational interaction. Specifically, she examined to what extent a speaker phonetically converges towards their partner's productions by comparing them to utterances obtained during a pre- and post-interaction reading task. Her findings suggested that phonetic convergence is driven by both the model talker's role, namely, giver or receiver, and the gender of the talker pair. Specifically, she observed greater convergence in men than in women and a significant interaction between role and gender indicating a different interpretation of role attributes in men and women. These differences have been attributed to the modulation of social contexts, communication goals and speaker roles.

Investigations into phonetic convergence have up to now focused on a limited range of languages, with a significant emphasis on English. Only a few studies have concentrated on French and they are primarily grounded in non-interactive paradigms (Delvaux & Soquet, 2007), non-conversational approaches (Lelong & Bailly, 2011) or specifically investigating regional phonological variation (Aubanel & Nguyen, 2010). Therefore, the aim of this study is to explore this phenomenon by contributing valuable data on conversational French, replicating Pardo's (2006) initial paradigm and extending analyses beyond the modulation by talker role and gender coupling in the pair, including a tentative exploration of phone combination in the target word (following Lelong & Bailly, 2011) and the influence of the total number of repetitions of that word within a given dyad (following Nielsen, 2011). We expect our results to align with Pardo's, finding evidence of increased convergence with the repeated item. Furthermore, we should observe a greater convergence effect among participants playing the role of givers compared to receivers, in male-male dyads compared to female-female dyads and in participants that have been exposed to a higher number of item repetitions. Lastly, in line with the findings of Lelong & Bailly (2011), we anticipate the observation of a phoneme-dependent effect, with mid-vowels giving rise to greater convergence compared to low or high ones.

Methods. In order to assess phonetic convergence in conversational interaction, we conducted a perceptual AXB task using speech data extracted from the Aix MapTask corpus (Bard *et al.*, 2013). This corpus was intentionally developed as an extension of the original HCRC Maptask (Anderson *et al.*, 1991), with a focus on fostering active engagement in collaborative social interactions among participants. Moreover, the task possessed the methodological advantage of eliciting spontaneous repetitions of identical items by both speakers. To build up the perceptual test, we extracted a subset of map task landmark labels that were reiterated by both members of a pair. Specifically, we selected four items - *blaireau*, *girafes, mirage, pyramide* - that were consistently repeated across all pairs of talkers. These repetitions were then employed as stimuli in the AXB task. Following the experimental paradigm designed by Pardo (2006), participants were instructed to assess similarity in pronunciation by choosing which of the A and B stimuli, as uttered by one participant, sounded more like the X stimulus, i.e., the same item but produced by their partner during the course of the interaction. In each trial three repetitions of a specific landmark label were presented, prompting participants to assess similarity by comparing the first and last items to the middle item. The X item was formed by a randomised instance of a target word uttered by either the giver or the receiver during the interaction, with the first and last occurrences excluded. The A and B items were formed by the corresponding partner's first and last occurrences of the same target

word. Following each response, the subsequent trial began 500 ms later. The presentation order of the triplets items was counterbalanced and the effects of timing, phones, talker role and pair sex were all tested.

Results. Following Kim *et al.*'s (2011) statistical analyses design, the preliminary data obtained on five participants (one male, all native French speakers with a mean age of 23 years old with no hearing impairment and a normal or corrected-to-normal vision) was fitted into a generalised linear mixed-effects model for binomial data (GLMM) on R, version 4.3.1 (R Core Team, 2023). Phonetic convergence, our response variable, was defined as the likelihood of choosing a last as opposed to a first repetition item during the AXB perceptual judgement task whereas the gender pairing of the dyad, the role of the model talker, the total number of repetitions of the target word within a given dyad and the phone combination, i.e. whether the target word contained a constraint of low and high vowels, like *mirage* did, or a combination of mid-vowels, as did *blaireau*, presented by that word were treated as either fixed-effect factors or covariates. Finally, to account for repeated measurements, the identity of the listener as well as that of the talker and their dyad were specified as random factors.

Overall, listeners seem to identify phonetic divergence rather than the expected convergence. More specifically, after model pruning, the total number of repetitions of a given target word within a single dyad during the map task seemed to yield a effect main effect on phonetic convergence rating (GLMM: pseudoR^2=0.021, χ^2=9.234, p=0.002 **), with a small albeit significant convergence pattern morphing into divergence as the number of repetitions increased. No significant effect of pair gender (i.e., male-male, female-female, or male-female; χ^2=5.851, p=0.054), role of the model talker (i.e., giver or follower; χ^2=2.835, p=0.092) or phone combination within the target word (χ^2=0.010, p=0.919) was found. Finally, a single association between the role endorsed by the model talker during the map task and phone combination was observed to significantly influence phonetic convergence (χ^2=8.488, p=0.004 **), with the presence of mid vowels within the target word leading to greater phonetic convergence solely when the model talker happened to be giving out the spatial instructions.

Discussion. These results, which seemingly contradict Nielsen's (2011) regarding the influence of repetitions on phonetic convergence, are to be carefully handled: not only are they based on a still preliminary exploration of data gathered on a very limited number of listeners, they only captured about 2,1% of the patterns observed, according to the computed determination coefficient pseudoR^2 - in other words, they call for a cautious reevaluation of the data, here under the form of a more thorough data collection, the results from which we hope to be able to present next May. Gathering a larger dataset would also allow us to better characterise a seemingly emerging trend regarding the influence of the gender pairing of the dyad on phonetic convergence, here to be found just above the threshold of significance level, according to which productions appeared to diverge in mixed and male-male pairings. If this were to be confirmed, our results would then stray away from those of Pardo (2006) and open up brand new reflexion axes regarding whether these results may simply be understood as by-products of our experimental paradigm, including elicitation task and convergence assessment design, and/or would be strategically driven by phonetic and/or socio-cultural specificities of the French language.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech* 34, 351–366.

Aubanel, V. & Nguyen, N. (2010). Automatic recognition of regional phonological variation in conversational interaction. *Speech Communication*. 52. 577-586. 10.1016/j.specom.2010.02.008.

Bard, E. G., Astésano, C., Turk, A., Nguyen, N., D'Imperio, M., Prévot, L., & Bigi, B. (2013). AIX MapTask : a (rather) new French resource for prosodic and discourse studies. *HAL (Le Centre pour la Communication Scientifique Directe)*.

Delvaux, V. & Soquet, A. (2007). The Influence of Ambient Speech on Adult Speech Productions through Unintentional Imitation. *Phonetica*. 64. 145-73. 10.1159/0000107914.

Fowler, C. A., Brown, J., Sabadini, L., and Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language* 49, 396–413.

Giles, H., et al., Speech accommodation theory: The first decade and beyond, in *Communication Yearbook*, M.L. McLaughlin, Editor. 1987, Sage Publishers: London, UK. p. 13-48.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105(2), 251–279.

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, 2(1).

Lelong, A. and G. Bailly (2011). Study of the phenomenon of phonetic convergence thanks to speech dominoes. *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issue*. A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud and A. Nijholt. Berlin, Springer Verlag, pp 280-293.

Namy, L. L., Nygaard, L. C., and Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology,* 21, 422–432.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, 39(2), 132-142.

Pardo JS. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America,* 119(4):2382-93. doi: 10.1121/1.2178720. PMID: 16642851.

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Science*, 27, 169–190.

Pickering, M. & Garrod, S. (2006). Alignment as the Basis for Successful Communication. *Research On Language And Computation*, 4. 203-228. 10.1007/s11168-006-9004-0.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Testing the generality of L1 phonetic precision as a predictor of L2 VOT acquisition

Marie K. Huffman[1], Katharina S. Schuhmann[2]

[1]Stony Brook University, United States of America
[2] Carl von Ossietzky Universität Oldenburg, Germany
Marie.Huffman@stonybrook.edu, Katharina.Schuhmann@uni-oldenburg.de

Introduction. Individual differences in second language pronunciation have many sources. One area of much recent interest is personal factors that may affect the structure of an individual's first and second language sound system, such as the consistency of their phonetic categories. Flege and Bohn (2021) in their revision of the Speech Learning Model SLM (as SLM-r) propose the Category Precision Hypothesis, which claims that category precision is a general endogenous trait and that individuals with more "precise" L1 phonetic categories may have an advantage in perceiving and thus producing second language sounds that differ from L1 sounds. While perceptual acuity is assumed to be the foundation of these effects, Flege and Bohn suggest that precision will also be evident in acoustic measures, so that speakers with more "precise" categories will show less acoustic variation (more compactness) in their phonetic categories. Several studies have in fact found relations between L1 category "compactness" and L2 pronunciation, as evidenced in acoustic data (e.g., Kartushina & Frauenfelder 2014; Kartushina et al. 2016; Huffman and Schuhmann 2021). Following this logic, then, a strong version of the Category Precision Hypothesis predicts that such effects should hold within individuals across different types of phonetic categories.

Huffman and Schuhmann (2021) reported that speakers with more compact L1 English voiceless (long lag) VOT categories early in a semester of college foreign language instruction had lower L2 Spanish voiceless stop VOT values at the end of the term. Here we consider additional data from these same individuals, to determine whether *multiple* L1 category compactness measures correlate with L2 Spanish voiceless VOT production. The L1 English properties we considered were compactness of the short lag VOT values for English [b] and [d], and formant and duration measures for L1 English vowels [i] and [ɑ]. If the compactness of multiple L1 phonetic categories correlates with L2 Spanish voiceless stop VOT, this would be strong support for category precision as an endogenous trait that influences L2 phonetics.

Methods. Participants were volunteers from the first three semesters of lower-level Spanish courses at a small liberal arts college on the eastern seaboard of the US. All reported having no extensive experience with any language other than English. Ten participants (9 F, 1M) were recorded as they read lists of English and Spanish words that contrasted in initial consonant and the following vowel. Target consonants were [p t b d] in word initial position after pause, followed by: [i] [e] or [a] for Spanish and [i], [eɪ] or [ɑ] for English. The target items in each language were pseudo-randomized with an equal number of filler items to form a list that was read three times by each subject. Subjects were recorded every two weeks from Week 2 to Week 12 of the semester. In Week 7, the speakers also participated in a short phonetic training session in which voicing differences between English and Spanish were illustrated and practiced (see Schuhmann and Huffman 2019 for details). Here, we compare measures of vowel and consonant category precision in L1 at the first recording (Week 2) with the mean voiceless stop VOT produced for L2 Spanish in the last recording (Week 12). VOT and vowel boundaries were segmented by hand in Praat (Boersma & Weenink 2023) by visual inspection of the waveform and spectrogram display. Praat scripts were used to automatically calculate VOT, vowel duration and formants at vowel midpoint. Formant values were then checked visually for errors and were remeasured when needed.

Results. English voiced stop short-lag VOT from Week 2 was analyzed for eight of our ten subjects. The other two subjects used prevoicing for English [b] and [d] to the extent that there were not enough short-lag tokens to allow a test of correlation with L2 VOT. For the eight speakers with sufficient short-lag data, we found that L2 Spanish voiceless stop VOT means at Week 12 showed a positive correlation with L1 English voiced stop short-lag VOT standard deviation (SD) in Week 2 (R^2 = .553, F(1,7) =7.41, p =.034). As shown in Figure 1 (blue dots), those with less variable L1 voiced stop short-lag (SL) VOT had lower Spanish voiceless stop mean VOT. Vowel formant variability was assessed by evaluating the standard deviation of F1 and F2 for L1 English target [i] and [ɑ], as well as vowel durations, at Week 2 for all ten participants. None of the vowel formant measures were found to correlate with L2 Spanish voiceless stop VOT means at Week 12 (F1 [ɑ] (R^2 = .261, F(1,7) =.585, p =.466); F2 [ɑ] (R^2 = .100, F(1,9) =.893, p =.372); F1 [i] (R^2 = .155, F(1,9) =1.47, p =.260); F2 [i] (R^2 = .024, F(1,9) =.197, p =.669). For illustrative purposes, we include the data from F2 standard deviation (SD) of L1 English [i] in Figure 1, plotted against Week 12 voiceless VOT means for L2 Spanish (red triangles). We also evaluated the durations of L1 English

[i] and [ɑ] in Week 2 L1 English to see if L1 vowel duration would correlate with Week 12 L2 Spanish voiceless VOT, but here too there was no correlation (R^2 = .216, F(1,7) =1.65, p =.245).

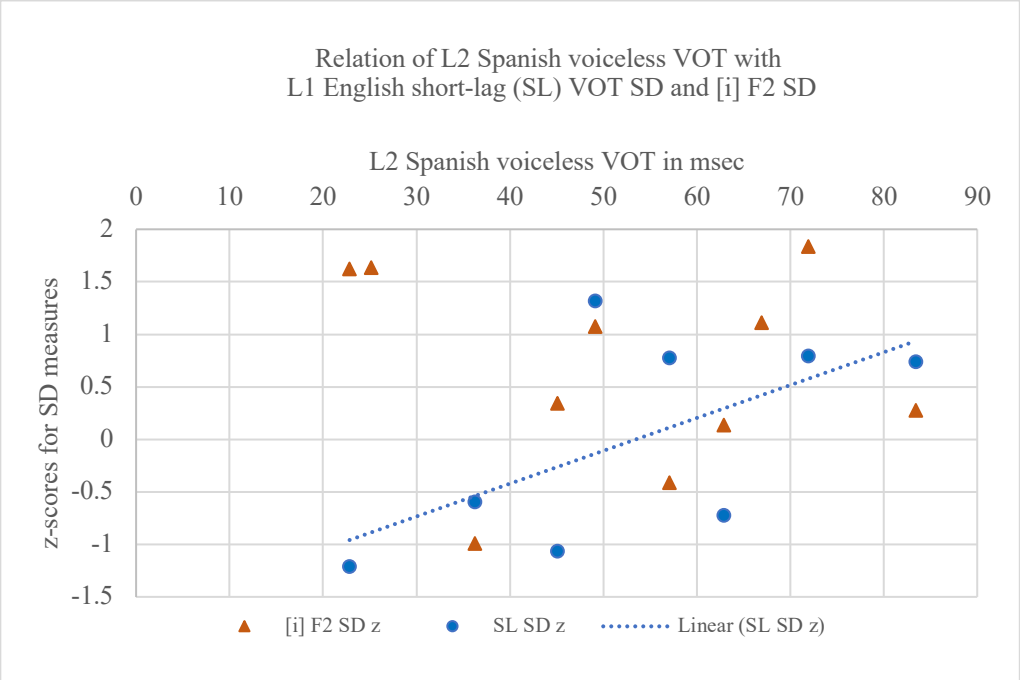


Figure 1. Correlation of L2 Spanish voiceless VOT (msec, x-axis) at Week 12 vs. L1 English voiced stop short-lag (SL) VOT SD and lack of correlation for [i] F2 SD at Week 2 (z-transformed, y-axis)

Discussion. We set out to examine the extent to which individual differences in (mean) L2 VOT may be related to the general compactness of a learner's L1 speech categories. We tested L2 Spanish mean voiceless stop VOT against compactness in L1 vowel production (formant and duration measures) and in short-lag VOT for English [b d]. Individual differences in L2 VOT were not associated with either type of vowel measurement. The vowel data presented here, taken alongside the short-lag and long-lag VOT data, do not support the claim that there is a *maximally general* personal precision factor that manifests in L1 and affects L2 learning progress. However, L2 VOT means were correlated with compactness in the L1 voiced stop short-lag VOT category. Thus, the L1 short-lag stop data, in combination with our previous results on L1 long-lag VOT data for the same subjects (Huffman and Schuhmann 2021), shows support for L1 category compactness as potentially playing a role in L2 phonetic learning. Overall, our data suggest that the way precision of L1 categories relates to L2 learning may vary for different phonetic properties within the same individual. Future phonetic analysis should examine how consistently L1 VOT category precision predicts L2 VOT across other L1:L2 pairings. More work is needed to test how consistently L1 VOT precision, as a predominantly temporal measure, correlates with other phonetic features that involve careful temporal coordination. Examining acquisition of both a new L2 VOT category and contrastive vowel or consonant length in the L2 should be particularly informative. Future research should also examine the extent to which perceptual acuity informs the cases when L1 production precision does not predict L2 production abilities.

References

Boersma, Paul & Weenink, David (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.02, retrieved 2 March 2023 from http://www.praat.org/

Flege, J. E., & Bohn, O.-S. (2021). The revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3-83). Cambridge: Cambridge University Press.

Huffman, M.K. & Schuhmann, K.S. (2021). The relation between L1 and L2 category compactness and L2 VOT learning. *Proc. Mtgs. Acoust.* 11 December 2020; 42 (1): 060011. DOI: 10.1121/2.0001421

Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, *5*, 1246. DOI: 10.3389/fpsyg.2014.01246

Kartushina, N., Hervais-Adelman, A , Frauenfelder, U. H., & Golestani, N. (2016). Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics,* 57, 21-39.

Schuhmann, K.S. & Huffman, M. K. (2019) Development of L2 stop voicing before and after pronunciation instruction *Journal of Second Language Pronunciation.* 03 December 2019 DOI: 10.1075/jslp.18018.sch

Malleability of speech sound representations: bite blocks and aftereffects

Xinyu Zhang[1], Ester Janse[1]

[1]Rabdoub University

With the acquisition of their first language, speakers acquire the mapping between articulator configurations and their consequent auditory and somatosensory outcomes. The question remains how malleable this motor-auditory mapping is in adult speakers. Namely, can speakers adapt to altered articulatory configurations and does the newly learned mapping linger after the manipulation is removed.

The current study is part of a larger project on the relationship between (representations in) sound production and perception where we investigate the effect of articulator manipulation on both speech acoustics and sound categorization in perception. More specifically, we focused on the effect of inhibiting tongue-height adjustment with a bite block, and the production and perception of vowels /ɪ/ and /ɛ/. Earlier, we found that sound categorization is not affected by having spoken with a bite block; i.e., neither the bite-block group, nor the control group changed their categorization of an /ɪ/ -to-/ɛ/ continuum from pre-test to post-test. To complement these sound categorization results, we compared the acoustics of two target vowels /ɪ/ and /ɛ/ contrasting in vowel height, produced by speakers who spoke normally, to those who spoke with a bite block inhibiting tongue-height adjustment. Additionally, to test the role of auditory feedback on the potential change in production, we compared productions of speakers who had access to their own bite block productions with those of speakers who had little auditory feedback of their own speech due to masking noise. We address the following research questions:

1) Are vowel acoustics (F1 and F2) of /ɪ/ and /ɛ/ affected by having a tongue-height-restricting bite block in the mouth during the speaking task as compared to baseline at pretest?
2) If so, does it matter whether speakers had access to their own obstructed productions?
3) Does having spoken with a bite block change subsequent vowel acoustics (i.e., after the bite block has been taken out, as compared to pretest production) for vowels /ɪ/ and /ɛ/?

Sixty healthy adult speakers were tested, 30 of which were assigned to the bite block group, and the other 30 to the no bite block control group. At test, within each group, half of the participants spoke with ordinary feedback during test block 1 followed by masking noise during test block 2. The other half of the group went through the reverse order. Participants were instructed to hold the bite block with the incisors at a set point and keep the tongue flat under the bite block. The production stimuli were bisyllabic pseudowords conforming to Dutch phonotactics (target vowels embedded in first, stressed syllable). Due to missing recordings, data of 54 speakers were analyzed.

Results show that having a bite block changed both the F1 and F2 (see Figure 1), be it in different ways (for F2) for the two vowels. For F1, this bite block effect was similar regardless of whether speakers had access to their own auditory feedback at test block 1 or not. For F2, the bite block effect is increased in the masking noise group, suggesting that auditory feedback does play a role in adapting to new articulator configurations (see Figure 2). In line with earlier literature, analyses showed no changes in adaptive speech behavior over trials. In terms of aftereffects, we found no difference between pretest and posttest in the F1 or F2 in either of the two vowels (see Figure 3). This implies that speakers reverted to their baseline production after the bite block was taken out, mirroring the lack of a perceptual change in categorization observed earlier, providing evidence for the stability of the stored representations.

Figure 1. F1-F2 scatterplots of each phoneme at pretest and test for the bite block group.

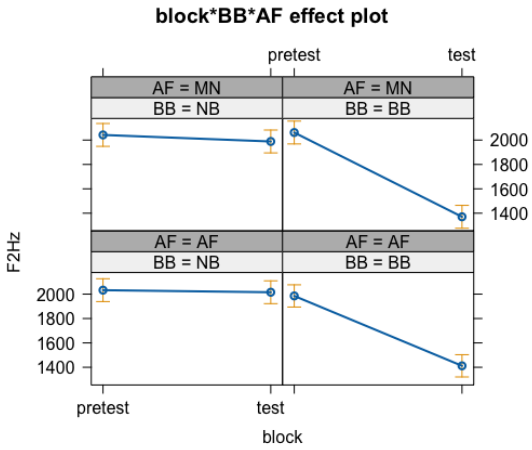


Figure 2. Effect plot of the interaction between bite block (NB: no bite block group, BB: bite block group), auditory feedback (AF: ordinary auditory feedback, MN: masking noise at test), and testing block (pretest vs test).

Figure 3. F1-F2 scatterplots of each phoneme at pre- and posttest for the group who had the bite block at test.

Explorations into speaker consistency in speech breathing and anticipation of upcoming phonetic content

Laura L. Koenig[1,2,3] *& Susanne Fuchs*[3]

[1]Adelphi University, USA
[2]Yale Child Study Center, USA
[3]Leibniz-Zentrum Allgemeine Sprachwissenschaft [ZAS], Germany
laura.koenig@yale.edu, fuchs@leibniz-zas.de

Introduction. Several previous studies have explored "pause postures" [e.g., Krivokapić *et al*., 2020] or "articulatory settings" [e.g., Gick *et al*., 2004], defined as oral placements unrelated to surrounding speech material. The term "posture" suggests a fairly stable position between speech events. Inter-speech pauses may or may not contain breath noise, which may be oral, nasal, or a combination of both (Lester & Hoit, 2014). When the mouth is open, one can assess whether breath noises show anticipatory influences of following speech. Some studies have reported such effects [Sarmah *et al.*, 2023; cf. Rasskazova *et al.*, 2019 for compatible kinematic results], but the extent to which phonetic context influences speech inhalation sounds remains a largely open question (Werner, 2023). Indeed, since breath noises can both be rather long, and precede speech by hundreds of milliseconds, this could reflect rather long-term anticipation, and one might expect that more coarticulatory effects would be observed when the inhalation is temporally closer to the speech. Finally, past work suggests that some aspects of quiet breathing and speech breathing may be consistent within an individual over days or even years (see Serré *et al.*, 2021 and citations therein). To our knowledge, speaker consistency has not been assessed for speech breath acoustics. In light of this past work, our current research questions are as follows: 1) Are speakers consistent over time in their use of oral and nasal inhalation patterns during speech? 2) Are they consistent in the stability of their breath acoustics across recording sessions? 3) To what extent do formants in breath noise correlate with upcoming speech sounds? 4) Do anticipatory effects depend on the duration between the breath noise and the speech?

Methods. We collected data from 6 female speakers of Northern German, ages 21–34. All speakers were recorded twice approximately 6 months apart as part of pilot explorations using an electro-optical stomatography system [Stone & Birkholz, 2020]. The current analysis assesses only the acoustic data, collected using a head-mounted microphone. In both recording sessions, speakers carried out a reading task and a more spontaneous speech task. For Study 2, the spontaneous task was a monologue about a favorite place or food to eat or prepare. For Study 1 it was an interactive game with the researcher in which interlocutors took turns "packing their suitcase" with an item at a time; in each turn, the speakers needed to list the items previously mentioned and add one.

In Praat (https://www.fon.hum.uva.nl/praat) version 6.2.05, breath noises were labelled. In total, 1776 breath noises were analyzed across the 6 speakers and 2 recording sessions. Breath noises were coded, based on audition, into oral, nasal, oral-to-nasal or nasal-to-oral breaths. A small number of breaths (N = 17/1776, <1% of the data) had more complex coding (e.g., simultaneous nasal+oral) and were subsequently excluded from the analysis. When the breath noise was perceived as involving mouth opening, we also labeled the onset and offset of the following vowel. Formants were obtained in the (oral) breath noises and the following vowel. Since breath durations could exceed 1 s in some cases, and the following syllable might be rather short, we assessed breath noise formants by obtaining averages in three 50 ms windows at the beginning, middle, and end of the breath, and vowel formants as an average over the first 30 ms. This allowed us to assess coarticulatory influences at the end of the breath noise, as well as change over time in the breath noise. Formant analysis in Praat used To Formant (burg), with a step size of 0.0025 s, 5 formants, maximum formant value = 5500 Hz, 0.025 window length, and pre-emphasis from 50 Hz. Vowel formants were reviewed and the N formants were adjusted to correct obvious mistakes. Such corrections were rather rare (c. 2% of vowels) and typically involved high back vowels. The following segment was categorized as being either a consonant or a vowel. Consonants were then separated according to place of articulation (removing uvular, where there was only one token) and vowels according to height, using a four-level classification.

Results. For question 1, we observe that the vast majority (84%, 1489/1776) of breaths were perceived to be oral. For a given speaker and recording session, oral breaths accounted for 53–98% of the tokens. Speakers were not necessarily consistent over time in their rates of oral breathing; as the most extreme example, sp3 went from 81% oral breathing in Study 1 to 53% in Study 2, where she used considerably more nasal (17%), nasal-to-oral (8%), and oral-to-nasal breathing (21%).

For question 2, the data show that formant ranges and averages of breath noise can vary considerably between recording sessions. In one speaker first and second formants changed in different directions across the two sessions.

For question 3, following Sarmah *et al.* (2023), we correlated formants in breath noises and the upcoming speech, considering place of articulation for consonants and height for vowels. The first formant of the breath noise was positively correlated with the upcoming speech for bilabial, labiodental, alveolar, and glottal places of articulation, but not velar or palatal (Figure 1, top). In those two places of articulation the slopes differed across speakers. When the data are separated by vowel height, the correlation slope is steeper for the open and close-mid vowels, and rather flat for the other two categories. Corresponding data for the second formant showed modest positive correlations between breathing and speech for all consonant places of articulation.

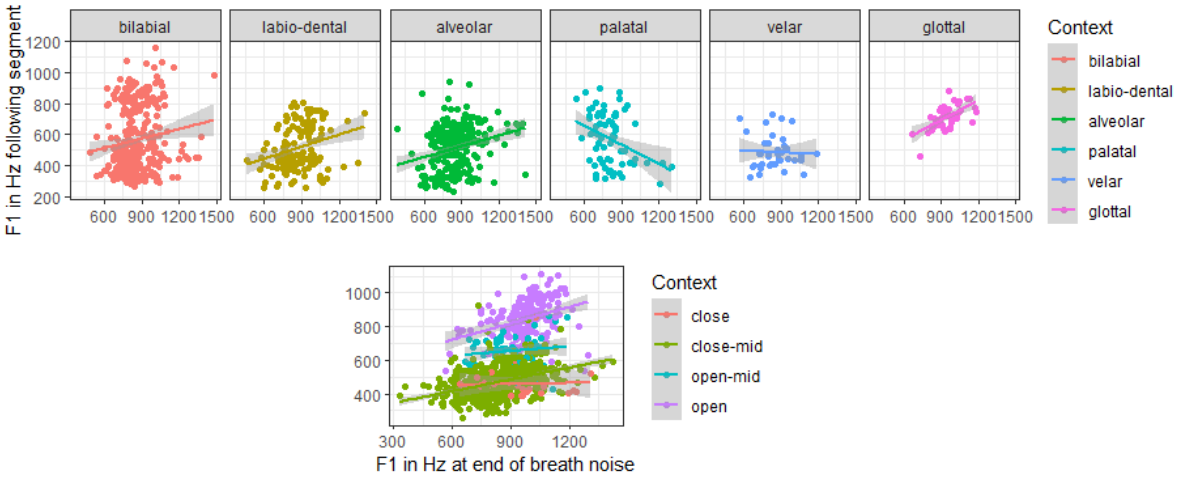


Figure 1: *Correlations between the first formant of the inhalation noise and the F1 of the following word, dividing out by consonant type (top) for consonant-initial words and vowel height (bottom) for vowel-initial words.*

Finally, for question 4, the distance in Hz between the first formant of the breath noise and upcoming syllable was generally flat as a function of duration between the breath and the speech, but tended more negative for palatal and velar places of articulation. For the second formant, relationships were again rather flat; exceptions were positive slopes for palatals and, to a lesser extent, bilabials.

Discussion. In contrast to Serré et al. (2023), who assessed kinematic trajectories of speech breathing, our acoustic results do not indicate that speakers are particularly consistent in their breath sounds, or oral-nasal inhalation patterns. Our auditory-only assessment of oral vs. nasal breathing may be less sensitive to the distinction between purely oral vs. oral+nasal inhalation as compared to Lester and Hoit (2014), who relied on nasal pressure and video recordings; however, as long as the mouth is open, formant data should provide information about oral postures during the inhalation. Whereas Sarmah et al. (2023) concluded that speech inhalation showed strong influences of upcoming phonetic context, our results show that the strength of this relationship depends on the nature of the upcoming consonant. The degree of anticipation, on the whole, is not strongly influenced by the duration between the breath and the speech.

References

Gick, B., Wilson, I., Koch, K., & Cook, C. (2004). Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica*, 61, 220–233.

Krivokapić, J., Styler, W., & Parrell, B. (2020). Pause postures: The relationship between articulation and cognitive processes during pauses. *Journal of Phonetics*, 79, 100953.

Lester, R. A., & Hoit, J. D. (2014). Nasal and oral inspiration during natural speech breathing. *Journal of Speech, Language, and Hearing Research*, 57(3), 734–742.

Rasskazova, O., Mooshammer, C., & Fuchs, S. (2019). Temporal coordination of articulatory and respiratory events prior to speech initiation. Proceedings of Interspeech 2019, 884–888.

Sarmah, P., Lalhminghlui, W., & Sharma, N. K. (2023). Sounds of <sil>ence: Acoustics of inhalation in read speech. *Proceedings of the International Conference on Speech and Computer*, pp. 314–321. Switzerland: Springer Nature.

Serré, H., Dohen, M., Fuchs, S., Gerber, S., & Rochet-Capellan, A. (2021). Speech breathing: Variable but individual over time and according to limb movements. *Annals of the New York Academy of Sciences,* 1505(1), 142–155.

Stone, S., & Birkholz, P. (2020). Articulation-to-speech using electro-optical stomatography and articulatory synthesis. *Proceedings of the 12th International Seminar on Speech Production*.

Werner, R. (2023). *The phonetics of speech breathing: pauses, physiology, acoustics, and perception*. Doctoral Dissertation, Universität des Saarlandes.

COMPARISON OF ACOUSTIC AND PHYSIOLOGICAL MEASURES OF COARTICULATION

Irfana, M[1]. & Fathima Nuha[2],

1. Assistant Professor in Speech Sciences, Department of Speech Language Pathology, All India Institute of Speech and Hearing, Mysore, India, email ID: irfana@aiishmysore.in **(Corresponding author)**
2. Speech Language Pathologist, All India Institute of Speech and Hearing, Manasagangothri, Mysore, India, email ID: nuha08651@gmail.com

Abstract

Introduction: Speech production is a complex process that involves the coordinated movement of different articulators, such as the tongue, lips, and jaw. The study of these movements can provide valuable insights into the nature of speech sounds and the mechanisms that underlie speech production. Coarticulation is one such measure that explains the influence of one sound on neighboring sound and the coordination of articulatory movements during speech production. It can be measured in different ways such as perceptual, acoustic and physiological methods. Acoustic and physiological measures explains the nature of speech production in better manner. Though studies have been done separately for acoustic and physiological methods to see the coarticulation, there is lack of comparison studies between the measures to verify the robustness of the same. Present study considered locus equation as acoustic measure and ultrasound imaging technique (UIT) as physiological method.

Aim and objectives: The aim of this study was to compare the acoustical and physiological coarticulatory measures across adult Malayalam speakers. Objectives of the study were (1) to compare the acoustical and physiological coarticulatory measures, (2) to analyze coarticulation across consonants, and (3) to check coarticulation across vowels.

Method: The study involved 14 adult native Malayalam speakers; none of them was having any history of speech or hearing disorders. Each participant was recorded with 18 different consonant-vowel (CV) combinations, where three different vowels (/a/, /i/, and /u/) and six different consonants (/k/, /g/, /t/, /d/, /ʈ/, and /ɖ/) were included as stimuli. Praat software was used to record and analyze acoustic measures and inbuilt Articulate Assistant Advanced software was used to record and analyze ultrasound tongue images for physiological parameters. The recordings were done simultaneously in a sound-treated room using a high-quality microphone.

Results: Findings of the study showed that acoustic measures of coarticulation such as goodness of fit and intersection were positively correlated with physiological measures. However, the correlation was not significantly different for all the places of articulation and vowel contexts. Coarticulation of retroflex was more than other consonants and vowel

/i/ had greater coarticulation than /a/ and /u/. There were no significant difference in coarticulation between voiced and unvoiced counter parts.

Discussion and conclusion: present study showed that locus equation is a robust acoustic measure, which had similar results as physiological study. Retroflex showed greater coarticulation which is in consensus with previous studies that showed higher complexity of tongue dynamics leads greater coarticulation and it exhibit influence of preceding and following phonemes. This can be applied to the vowel /i/ that had greater coarticulation than /a/ and /u/.

Keywords: Acoustics, physiology, ultrasound, locus equation, frequency, tongue

Simulation-based Bayesian inference of state feedback control model parameters to fit f_0 perturbation responses in laryngeal dystonia

Jessica L. Gaines[1], Kwang S. Kim[2], Ben Parrell[3], Vikram Ramanarayanan[4,5], Alvincé L. Pongos[1], Srikantan S. Nagarajan[4], John F. Houde[4]

[1]UC Berkeley – UCSF Graduate Program in Bioengineering, University of California-San Francisco, San Francisco, California, USA
[2]Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, Indiana, USA
[3]Department of Communication Sciences and Disorders, University of Wisconsin-Madison, Madison, Wisconsin, USA
[4]Department of Otolaryngology, University of California-San Francisco, San Francisco, California, USA
[5]Modality.ai, San Francisco, California, USA

jessica.gaines@berkeley.edu, kwangkim@purdue.edu, bparrell@wisc.edu, vikram.ramanarayanan@modality.ai, alvince_pongos@berkeley.edu, srikantan.nagarajan@ucsf.edu, john.houde@ucsf.edu

Introduction. Experiments observing participants' behavioral response to a mid-utterance shift in voice fundamental frequency (f_0) have been widely used to study the impacts of various neurological conditions on the motor control of pitch (e.g., Kothare et al., 2022). However, when differences between a clinical group and a control group are observed in f_0 perturbation response, it can be difficult to ascribe which neural mechanisms may have led to these differences. For example, compared with individuals with healthy voice, those with laryngeal dystonia (LD) have shown slightly reduced magnitude and increased oscillations in response to an f_0 perturbation compared to controls (Kothare et al., 2022). Computational modeling can help address this challenge by simulating how the behavioral response might change as a result of changes in interpretable model parameters. By fitting model parameters to optimally simulate an observed behavioral response, we can generate hypotheses about how two groups may differ in terms of these parameters.

Methods. In this investigation, Bayesian inference was used to fit five tunable parameters of a state feedback control (SFC) model of voice fundamental frequency to the group average f_0 perturbation response of the LD group and the control group (Kothare et al., 2022). In this model (Houde & Nagarajan, 2011; Houde et al., 2014), a controller generates motor commands based on an internal estimate of the laryngeal state (position and velocity). An efference copy of these commands is used to predict the subsequent laryngeal state, and thus the expected sensory consequences of the movement. The plant, a simplified model of the larynx (a state space representation of a spring-mass system) generates simulated auditory and somatosensory feedback. This simulated feedback, which is delayed and combined with Gaussian noise, is compared with the expected sensory feedback and then the error between these two signals is used to update the internal estimate of laryngeal state. This update is weighted by a Kalman gain that scales the weight of the feedback based on the amount of noise in the feedback. The tunable parameters affecting the model output are auditory delay (Δ_a), somatosensory delay (Δ_s), auditory feedback noise covariance (σ_a), somatosensory feedback noise covariance (σ_s), and controller gain (g_c). The noise variables were transformed to the alternative parameters overall feedback noise covariance σ and feedback noise ratio r such that $\sigma_a=\sigma$ and $\sigma_s=\sigma/r$ in order to separate the effects of absolute level of feedback noise from the relative level of noise between sensory modalities. Thus the tunable parameter set was $\{\Delta_a, \Delta_s, \sigma, r, g_c\}$ (Gaines et al., in prep).

The sbi package in Python (Cranmer et al., 2020; Tejero-Cantero et al., 2020) was used to infer likelihood distributions across values for each parameter for both the LD group and control group. First a simulator (here the SFC model) was used to generate a set of parameter sets and their resulting model outputs (here the f_0 perturbation response). Using this data, a neural network was trained to predict the posterior likelihood of a parameter set given an empirical observation analogous to the model output. The posterior was then sampled to create a likelihood distribution for each

parameter. Glass's delta was used to quantify the effect size of the group difference for each parameter. Bootstrapping the posterior was used to calculate a standard error on the effect size. Finally, to verify the quality of the model fit, the median value of each distribution was used as input to the SFC model and the output was overlaid with the behavioral data. The quality of the fit was quantified using the root mean square error (RMSE) between each behavioral response and the corresponding simulated response.

Results. As seen in **Figure 1**, the parameter that is most different between the LD group and the control group is the feedback noise ratio parameter, with an effect size of 17.32 ± 0.05. The controller gain parameter has a moderately large effect size and the likelihood distributions for all other parameters are largely similar between the two groups, suggesting that the differences in pitch perturbation response can be mostly ascribed to the relative noise between sensory modalities. However, while the simulated response for the control group is very similar to the behavioral response (RMSE = 0.52 cents or 4.7% of the total magnitude of the response), the simulated response for the LD group aligns less well with the corresponding behavioral data (RMSE = 1.77 cents or 13.9% of the total magnitude of the response).

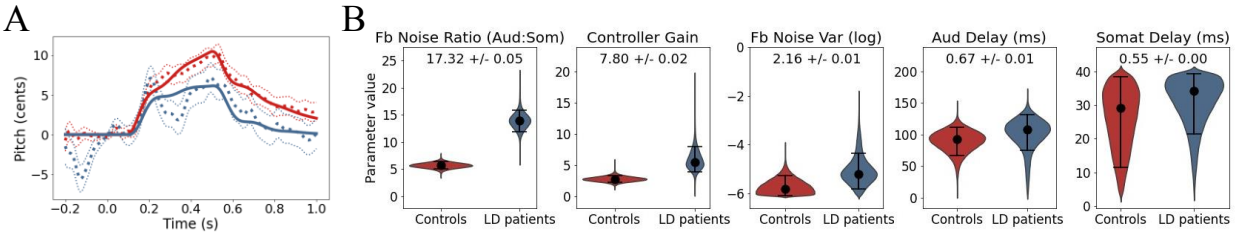


Figure 1: A) Behavioral f_0 response from Kothare et al. (2022; dotted lines) with SFC model output from the fit parameters (solid lines) for LD (blue) and control (red). B) Likelihood distributions for each parameter.

Discussion. The parameter with the greatest difference in value between the LD group and the control group is the feedback noise ratio parameter, which describes the relative amount of noise between auditory and somatosensory modalities. In the SFC model, auditory and somatosensory feedback noise are used to calculate Kalman gain. This result suggests that in the LD group, a greater ratio of auditory to somatosensory feedback noise compared with the control group leads to a lower weighting of auditory feedback, which causes a lower magnitude response to an auditory error. Interestingly, the feedback noise variance parameter σ, which adjusts the absolute level of noise for both sensory modalities, had a much smaller effect size, showing that the absolute amount of auditory and somatosensory feedback noise is less important to group differences than how these values compare to each other.

The behavioral data of the LD group shows a few large oscillations that the model was unable to capture, leading to a lower quality of fit for this data than for that of the control group. This may indicate additional group differences that simple changes in parameter values could not explain, such as a difference in the architecture of the neural systems controlling f_0, or a difference in laryngeal dynamics that could not be captured by our simplified larynx model. Future work will investigate how the SFC model may simulate these oscillations.

References

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences 117*(48), 30055-62. doi: 10.1073/pnas.1912789117.

Gaines, J.L., Kim, K.S., Parrell, B., Ramanarayanan, V., Pongos, A.L., Nagarajan, S.S., & Houde, J.F. (in prep.). Bayesian inference of state feedback control parameters for f_0 perturbation responses in cerebellar ataxia.

Houde, J.F. & Nagarajan, S.S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience 5*, Article 82. doi: 10.3389/fnhum.2011.00082

Houde, J.F., Niziolek, C.A., Kort, N., Agnew, Z., & Nagarajan, S.S. (2014, May 5-8). Simulating a state feedback model of speaking. 10th International Seminar on Speech Production, Cologne, Germany.

Kothare, H., Schneider, S., Mizuiri, D., Hinkley, L,, Bhutada, A., Ranasinghe, K., et al. (2022). Temporal specificity of abnormal oscillations during phonatory events in laryngeal dystonia. *Brain Communications 4*(2), fcac031. doi: 10.1093/braincomms/fcac031.

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Goncalves, P.J., et al. (2020). Sbi: A toolkit for simulation-based inference. *Journal of Open Source Software 5*(52), 2505. doi: 10.21105/joss.02505.

Analysis of tongue movement (quasi)-"steady-states" using General Tau theory

Alice Turk[1], Benjamin Elie[1], Cedric Macmartin[1], David N. Lee[1]

[1]The University of Edinburgh

a.turk@ed.ac.uk, b.elie@ed.ac.uk, c.macmartin@ed.ac.uk, d.n.lee@ed.ac.uk

Introduction. A challenging issue for the analysis of speech articulatory trajectories are regions of very low-velocity movement. Should these regions of low velocity be treated as components of movements towards and away from a target at a single time point (e.g. at a velocity zero or at a tangential velocity minimum)? Or should these low velocity regions be treated as (quasi) "holds" or "steady-states"? These theoretical alternatives have implications for studies which rely on identifying movement "onsets" and "offsets", including studies of movement coordination, and movement duration.

One approach to this problem has been to use a velocity thresholding method, (e.g. Katsika et al. 2014), in which movement onsets and offsets are diagnosed at a fixed threshold of the movement's peak velocity (typically 20%). However, it is unclear whether this diagnostic is appropriate when movements are compared that have different peak velocities.

In this presentation, we explore an alternative approach. We address the (quasi) "steady state" issue using Lee's General Tau theory (Lee 1998) approach to modelling speech movement trajectories, shown to out-perform several other models (Elie *et al.* 2023). Tau theory assumes that purposeful movements close gaps from a current state to a target state, where gaps can be distance gaps, angle gaps, etc. Time-varying tau is the time-to-gap closure (or time to movement target achievement) at the current movement rate. Tau theory assumes that each planned speech movement trajectory X couples onto an intrinsic "Tau Guide" τ_G: $(\tau_X = k\tau_G)$; $\tau_G(t) = \frac{1}{2}\,(t - \frac{T^2}{t})$; where T = movement duration, and t runs from 0 to T. The coupling constant k determines the shape of the velocity profile. The gap is guaranteed to close at the prescribed time as long as coupling occurs before the end of the movement.

Methods. Our study consists of Tau theory analyses of tongue body movement data in 350+ post-pausal 'yeah' tokens in the ESPF Doubletalk corpus (Scobbie et al. 2013, Turk et al. 2013). Movements are analysed from zero-to- zero velocity in a single dimension determined via Principal Component Analysis, from a starting position to a position close to the hard palate for /j/ (the "up" movement segment), and from this position to a lower tongue position for /E/, (the "down" movement segment). Zero-to-zero velocity movements shorter than 50 ms are not analysed. Best tau fits are determined via an algorithm which finds the parameter values of the tau curve which minimizes the RMSE discrepancy, normalized by movement amplitude, between the observed trajectory and the candidate tau fit. Our strategy for answering our research question for each movement, will be to 1) compare the best tau fit(s) assuming no (quasi)-steady state (i.e. from zero-to-zero velocity), to 2) the best tau fit(s) assuming a steady state. These comparisons will be assessed by measuring the average fit errors (RMSE normalized by movement amplitude), the percentage of the duration of the zero-to-zero velocity segment that corresponds to the "steady state" case, as well as the percentage of variance accounted for by the PCA analysis. In addition, we will evaluate best fits with and without "steady states" while allowing for points at the beginning of each fit to be "dropped" from the error analysis (cf. Lee & Schögler 2009).

Results. Results from analyses assuming no steady state show that the median fit error is 3.5% for Up movements (s.d. 2.7%), and 2.9% for Down movements (s.d. 2%). We are currently exploring appropriate algorithms for diagnosing fits when steady states are assumed. In our first attempt, we relaxed the constraint that the fit should run from 0 to 0 velocity, and iteratively searched for the best tau fits whose onsets and endpoints are within +/- 15 ms windows of the previous best tau fit to find the new best tau fit. Figure 1 shows examples from two tokens of *yeah*, 1 row each, for fits running from 0 to 0 velocity (left panels), as well as the fits when the 0 to 0 velocity constraint is relaxed (right panels). For the top example, when the fits run from 0 to 0 velocity, the fit errors are 3.9% for the Up movement, and 3.1% for the Down, with k_{Up} = .499 and k_{Down}= .424. When the zero-to-zero velocity constraint is relaxed, the fits are improved (new fit errors are .5% for the Up movement, and .6% for the Down movement, k_{Up} = .459; k_{Down}= .325). Note that a small steady state can be seen in white at the top of the curve.

The example on the bottom row of Figure 1 shows fit errors for the zero-to-zero velocity case (bottom left) of 8.7% for the Up movement, and 1.9% for the Down movement, with k_{Up}= .92, and k_{Down}= .473. When the zero-to-zero velocity constraint is relaxed (bottom right), the errors again improve (.39% for the Up movement, and .18% for the Down movement; k_{Up} = .399; k_{Down}= .980). However, this example shows that the best fit for the Down movement (red) after relaxing the 0 to 0 velocity constraint ends inappropriately during a high velocity portion of the trajectory.

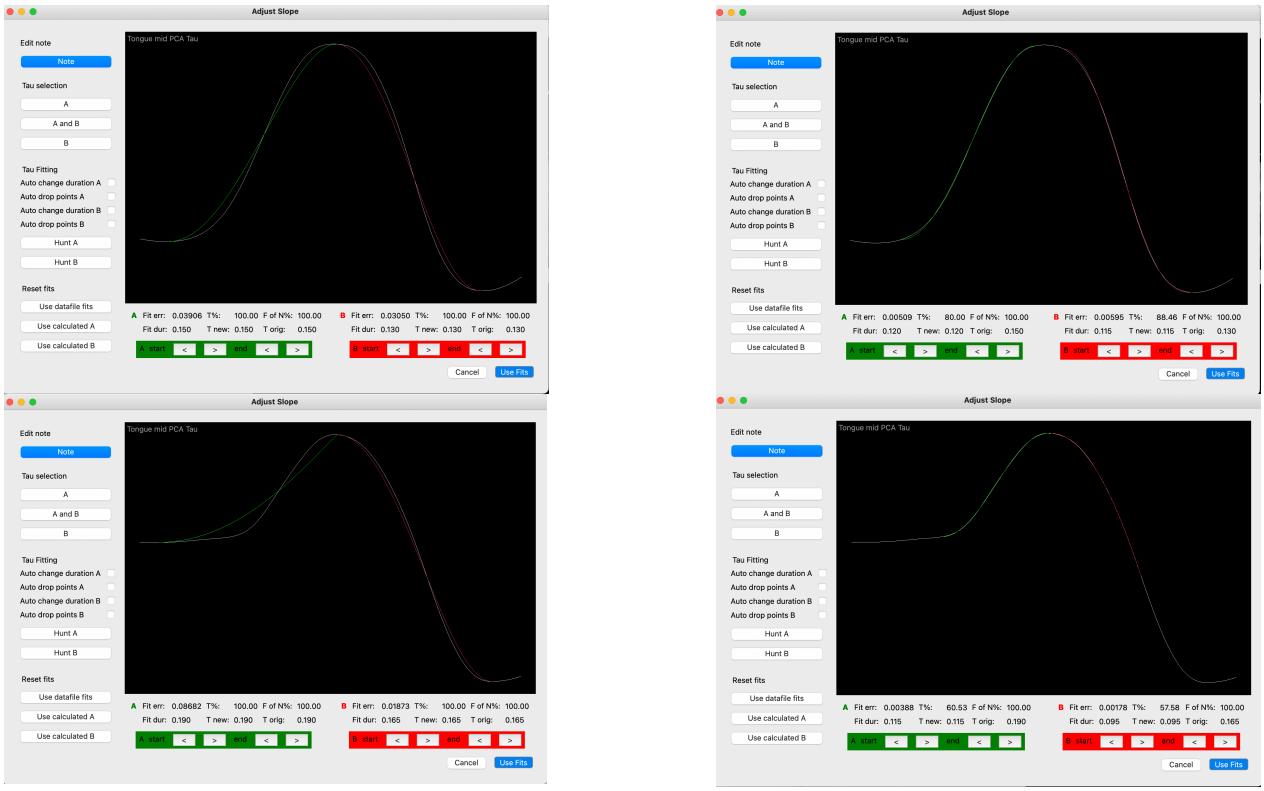


Figure 1: *Screenshot views of two examples (1 per row) from our EmaTViewer software which show Left panels: Best tau fits from 0 to 0 velocity (in green and red); and Right panels: Best tau fits when the 0 to 0 velocity constraint was relaxed; along with the fit errors (Fit err) and fit durations (Fit dur).*

Discussion. Results for cases such as the one shown in the bottom panels suggest that the "steady state" algorithm should be constrained so that the steady states do not include high velocity parts of movement. We will therefore modify our algorithm so that steady states are constrained to not include the movement acceleration maximum or deceleration minimum.

In addition, we plan to present tau fit analyses with and without steady states for analyses of movement segments based on the tangential distance curve, that run from tangential velocity minimum to tangential velocity minimum.

References

Elie, B., Lee, D. N., & Turk, A. (2023). Modeling trajectories of human speech articulators using general Tau theory *Speech Communication*, *151*, 24-38.

Katsika, A., Krivokapić, J., Mooshammer, C., Tiede, M., & Goldstein, L. (2014). The coordination of boundary tones and their interaction with prominence. *Journal of Phonetics*, *44*, 62-82.

Lee, D. N. (1998). Guiding movement by coupling taus. *Ecological Psychology*, *10*(3-4), 221-250.

Lee, D. N., & Schögler, B. (2009). Tau in musical expression. In S. Malloch & C. Trevarthen (Eds.), *Communicative musicality: Exploring the basis of human companionship* (pp. 83-104).

Scobbie, J. M., Turk, A., Geng, C., King, S., Lickley, R., & Richmond, K. (2013). The Edinburgh Speech Production Facility Doubletalk corpus. *Interspeech 2013*, 764-766.

Turk, A., Scobbie, J. M., Geng, C., et al. (2022). *Edinburgh Speech Production Facility (ESPF) Doubletalk* corpus. Edinburgh DataShare. https://doi.org/https://doi.org/10.7488/ds/3508

.

Brain changes associated with stuttering therapy and spontaneous recovery

Nicole E. Neef[1], Soo-Eun Chang[2]

[1]Department of Diagnostic and Interventional Neuroradiology, University Medical Center Göttingen, Germany
[2]Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA
nicole.neef@med.uni-goettingen.de, sooeunc@med.umich.edu

Introduction. Stuttering is a developmental speech disorder characterized by involuntary interruptions in the normal flow of speech. In adulthood, therapy for stuttering can be effective in improving fluency, while in childhood, individuals often spontaneously recover. Nearly three decades of neuroimaging research has provided unprecedented insights into the brain's signatures of developmental stuttering. While neuroimaging is not yet a tool to diagnose and clarify treatment for developmental stuttering on an individual basis, imaging findings have nevertheless (1) caused a paradigm shift towards classifying stuttering as a neurodevelopmental speech disorder, (2) elucidated neural signatures of stuttering as a trait, i.e. permanent cerebral characteristics, and as a state, i.e. a transiently occurring functional impairment, (3) detected markers of persistency and spontaneous recovery, (4) provided neurobiologically based explanations for treatment effects, and (5) motivated novel therapeutic strategies (Neumann & Neef, in press). Here we briefly summarize neuroimaging findings on neuroplasticity in stuttering from studies of adults and children who stutter relevant to significantly reduced stuttering (Neef & Chang, 2024).

Methods. Our qualitative summary was based on 13 neuroimaging studies with adults who underwent short-term (Lu et al. 2012; Toyomura et al. 2015; Lu et al. 2017) or long-term speech restructuring therapy (De Nil et al. 2001; De Nil et al. 2004; Kell et al. 2009; Kell et al. 2018; Neumann et al. 2018; Korzeczek et al. 2021; Neef et al. 2021; Neef et al. 2022), pharmacological intervention (Maguire et al., 2021) or speech training combined with neurostimulation (Chesters et al. 2021). Furthermore, we converge findings from three longitudinal studies on a cohort of children who stutter and who recovered from stuttering (Chow & Chang 2017; Garnett et al. 2018; Chow et al. 2023).

Results. Interventional studies in adults, although few in number, and in sample size and statistical power, have revealed patterns of potential functional reorganization within and beyond the speech network (Figure 1, right panel). Functional neuroimaging revealed neuroplasticity potential in brain regions and circuits that support speech-related auditory-motor processes and speech motor learning such as the cerebellum, cortico-basal ganglia circuits, and cortico-cortical circuits including the dorsal motor cortex, inferior frontal gyrus, insula, supplementary motor area, supramarginal gyrus, and posterior superior temporal gyrus. Unlike brain activity, brain structures showed no therapy-associated changes, neither in grey nor in white matter structures. Different from therapy-induced changes in adults, children who spontaneously recovered from stuttering showed primarily an age-related growth in white matter structures that enable fast and accurate sequential speech movements. These white matter structures, including corticospinal tract, superior longitudinal fasciculus, arcuate fasciculus, somatomotor part of the corpus callosum and cerebellar peduncles (Figure 1 left panel), interconnect gray matter regions that were found to show significant reductions in volume for children with persistent stuttering, including the left ventral motor cortex and the left dorsal premotor cortex. Spontaneous recovery was in addition linked to left ventral premotor cortex volume measures that were intermediate between those of children who do not stutter (controls) and persistent children who stutter, and less gyrification in premotor medial areas with age, including the presupplementary motor area and the supplementary motor area. Accordingly, spontaneous recovery implicates brain circuits involved in speech initiation, processing the metrical structure of the speech motor plan, and sensory feedback control. Remarkably, children who recover also showed neuroplasticity in grey matter structures associated with speech motor learning and feedforward control.

Discussion. Treatment-induced improvement of overt stuttering during adulthood requires extensive training and resources. Spontaneous recovery in children on the other hand, is relatively common, reported to be upwards of 80%, and results in complete alleviation of symptoms with no effort or internal struggle to produce fluent speech. It is undisputed that the brain has a higher potential for neuroplasticity during childhood than in adulthood. Our qualitative synthesis reflects a high potential for therapy-associated *functional* reorganization in speech-related cortical and subcortical regions in adults who stutter and a high potential for *structural* reorganization in speech-related grey and white matter regions in children who spontaneously recover. In summary, our review highlights theories and models of neurofunctional reorganization of speech fluency and motivates future studies on the potential of using non-invasive brain stimulation to improve treatment efficacy in individuals who stutter who wish to work on improving fluency.

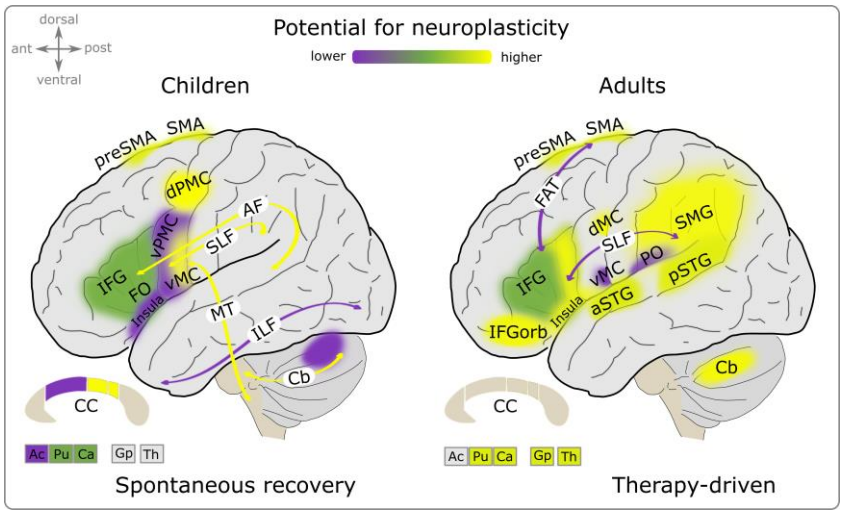


Figure 1: *Brain correlates for structural reorganization associated with spontaneous recovery and functional reorganization associated with therapy-induced improvements in stuttering (Neef & Chang, 2024).*

Abbreviations: Ac, nucleus accumbens; AF, arcuate fasciculus; aSTG, anterior superior temporal gyrus; Ca, caudate nucleus; Cb, cerebellum; CC, corpus callosum; dMC, dorsal primary motor cortex; dPMC, dorsal premotor cortex; FAT, frontal aslant tract, FO, frontal operculum; Gp, globus pallidus; IFG, inferior frontal gyrus; ILF, inferior longitudinal fasciculus; IFGorp, inferior frontal gyrus pars orbitalis; inferior longitudinal fasciculus; MT, motor tracts; pSTG, posterior superior temporal gyrus; PO, parietal operculum; Pu, putamen; SLF, superior longitudinal fasciculus; SMA, supplementary motor area, SMG, supramarginal gyrus; Th, thalamus; vMC, ventral primary motor cortex; vPMC, ventral premotor cortex.

References

Chesters, J., Möttönen, R. & Watkins, K. (2021). Neural changes after training with transcranial direct current stimulation to increase speech fluency in adults who stutter. doi:10.31219/osf.io/8st3j.

Chow, H. M. & Chang, S. (2017). White matter developmental trajectories associated with persistence and recovery of childhood stuttering. *Human Brain Mapping* 38(7), 3345–3359.

Chow, H. M., Garnett, E. O., Koenraads, S. P. C. & Chang, S. (2023). Brain developmental trajectories associated with childhood stuttering persistence and recovery. *Developmental Cognitive Neuroscience* 60, 101224.

Garnett, E. O., Chow, H. M., Nieto-Castañón, A., Tourville, J. A., Guenther, F. H., & Chang, S. (2018). Anomalous morphology in left hemisphere motor and premotor cortex of children who stutter. *Brain* 141(9), 2670–2684.

Kell, C. A., Neumann, K., Behrens, M., Wolff von Gudenberg, A. & Giraud, A. (2018). Speaking-related changes in cortical functional connectivity associated with assisted and spontaneous recovery from developmental stuttering. *Journal of Fluency Disorders* 55, 135–144.

Kell, C. A., Neumann, K., von Kriegstein, K., Posenenske, C., Wolff von Gudenberg, A., Euler, H. A. & Giraud, L. (2009). How the brain repairs stuttering. *Brain* 132(10), 2747–2760.

Korzeczek, A., Primaßin, A., Wolff von Gudenberg, A., Dechent, P., Paulus, W., Sommer, M. & Neef, N. E. (2021). Fluency shaping increases integration of the command-to-execution and the auditory-to-motor pathways in persistent developmental stuttering. *NeuroImage* 245, 118736.

Lu, C., Chen, C., Peng, D., You, W., Zhang, X., Ding, G., Deng, X., Yan, Q. & Howell, P. (2012). Neural anomaly and reorganization in speakers who stutter. *Neurology* 79(7), 625–632.

Lu, C., Zheng, L., Long, Y., Yan, Q., Ding, G., Liu, L., Peng, D. & Howell, P. (2017). Reorganization of brain function after a short-term behavioral intervention for stuttering. *Brain and Language* 168, 12–22.

Maguire, G. A., Yoo, B. R. & SheikhBahaei, S. (2021). Investigation of Risperidone Treatment associated with enhanced brain activity in patients who stutter. *Frontiers in Neuroscience* 15, 598949. doi:10.3389/fnins.2021.598949.

Neef, N. E., and Chang, S.-E. (2024). Knowns and unknowns about the neurobiology of stuttering. PLOS Biol. 22, e3002492. doi: 10.1371/journal.pbio.3002492.

Neef, N. E., Korzeczek, A., Primaßin, A., Wolff von Gudenberg, A., Dechent, P., Riedel, C. H., Paulus, W. & Sommer, M. (2022). White matter tract strength correlates with therapy outcome in persistent developmental stuttering. *Human Brain Mapping* 43(11), 3357–3374.

Neef, N. E., Primaßin, A., Wolff von Gudenberg, A., Dechent, P., Riedel, H. C., Paulus, W., & Sommer, M. (2021). Two cortical representations of voice control are differentially involved in speech fluency. *Brain Communications* 3(2), fcaa232-. doi:10.1093/braincomms/fcaa232.

Neumann, K., Euler, H. A., Kob, M., Wolff von Gudenberg, A., Giraud, A., Weissgerber, T. & Kell, C. A. (2018). Assisted and unassisted recession of functional anomalies associated with dysprosody in adults who stutter. *Journal of Fluency Disorders* 55, 120–134.

Neumann, K. & Neef, N. E. (in press). Neuroimaging findings in stuttering. In A. am Zehnhoff-Dinnesen, J. Sopko, M. Monfrais-Pfauwadel, K. Neumann (Eds.), Phoniatrics II Speech and Speech Fluency Disorders – Literacy Development Disorders. Springer Nature.

De Nil, L. F., Kroll, R. M. & Houle, S. (2001). Functional neuroimaging of cerebellar activation during single word reading and verb generation in stuttering and nonstuttering adults. *Neuroscience Letters* 302(2–3), 77–80.

De Nil, L. F., Kroll, R. M., Lafaille, S. J. & Houle, S. (2004). A positron emission tomography study of short- and long-term treatment effects on functional brain activation in adults who stutter. *Journal of Fluency Disorders* 28(4), 357–380.

Toyomura, A., Fujii, T. & Kuriki, S. (2015). Effect of an 8-week practice of externally triggered speech on basal ganglia activity of stuttering and fluent speakers. *NeuroImage* 109, 458–468.

Segmental durations and the vowel length contrast in fast speech in Hungarian

Andrea Deme[1], Kornélia Juhász[1,2], Zsuzsa Szánthó[1], Szabina Zsoldos[1], Reinhold Greisbach[3]

[1]ELTE Eötvös Loránd University, Budapest, Hungary
[2]HUN-REN Hungarian Research Centre for Linguistics, Budapest, Hungary
[3]University of Cologne, Cologne, Germany

deme.andrea@btk.elte.hu, juhasz.kornelia@nytud.hun-ren.hu,
zsuzsa.szantho@gmail.com, szamboc@gmail.com, reinhold.greisbach@uni-koeln.de

Introduction. Fast speech is the result of speech sounds produced shorter. However, it is expected that in terms of duration, not each segment may be reduced to the same extent in fast speech. Due to their homogenous structure throughout the total segmental duration, and the lack of an obstruction in the oral cavity, vowels are expected to be more flexible in this sense than (prototypical) consonants (i.e., obstruents) (Kozhevnikov & Chistovich 1965; Wood 1973; Gósy 2004; Lo & Sóskuthy 2023), which feature an obstruction in the mouth and can have a complex inner structure. Furthermore, differences are also expected according to phonemic length of the segments: in Japanese, it was shown that long vowels are affected more by speech rate than short vowels (i.e., they are more reduced or lengthened in fast and/or slow speech, respectively) (Hirata 2004), thus duration differences of long vs. short pairs reduced, but duration ratios were maintained in fast speech. In Korean, however, all vowels were reduced to a similar extent in fast speech, and no asymmetries were found (Magen & Blumstein 1993). In these languages, phonemic length contrast in vowels is expressed primarily by durational differences.

In Hungarian, vowel length is also phonologically distinctive. Traditionally, it is assumed that this opposition is implemented phonetically as a durational and spectral difference in open/low vowels (/ɛ/ vs. /eː/ and /ɒ/ vs. /aː/), but in more close/higher vowels (e.g., /i/ vs. /iː/; /u/ vs. /uː/), only durational differences can be found (Gósy 2004). With respect to the effect of speech rate on vowel and consonant durations, and the duration of phonologically long and short vowels at fast and normal/comfortable speech rates in Hungarian, we find no replicable and/or systematic analyses. However, we have recurrent evidence that the above outlined durational asymmetries are at work, that is, i) consonants are more resistant to speech rate effects (i.e., vowels reduce more in fast speech than consonants) (Magdics 1969), and ii) long vowels are affected more by speech rate than short vowels (i.e., long vowels reduce to a higher degree than short vowels) (Magdics 1969; Mády 2008). In the present study, we investigated these two hypotheses in acoustic data obtained in real words. Additionally, iii) we examined if the difference and the ratio of long and short vowels' duration is maintained across different speech rates, and tested if phonological length opposition is invariant as a function of tempo.

Methods. We analyzed CVC shaped real words in the production of 15 Hungarian speaking females. In these sequences, V was one of the following 6 vowels that constitute long-short vowel pairs in Hungarian: /u/, /uː/, /i/, /iː/, /ɒ/, or /aː/. In the onset, we placed laryngeal or alveolar consonants: /h z s t r/. In the coda, velar and alveolar consonants were positioned: /z t d k n r/. As a result, target sequences did not constitute minimal pairs, hence probably did not facilitate exaggeration of contrastive features of segments (e.g., vowel length). Speakers produced target words in carrier sentences, where the target word bore sentence level accent: *Legyen* <target word>! 'Let it be <target word>!' We recorded samples in two speech rate conditions: at i) comfortable speech rate ("normal" speech), and ii) maximum speech rate ("fast" speech). Maximum speech rate was achieved by the method of Greisbach (1992): speakers repeated each target sentence several times starting with a comfortable tempo (marked as normal speech later on in the analysis), and then, they started to repeat the same item several times trying to speak faster and faster at each repetition (until articulation broke down or speakers ran out of air). Each participant produced 6 of these sets (i.e., one normal rate variant followed by fast repetition variants) for each target word resulting in 72 sets (144 test tokens) per speaker in total. We labeled all sets manually in Praat (Boersma & Weenink 2022): we segmented each word, checked their durations, and labeled the shortest repetition as the fast speech variant, while we always took the first item produced at a comfortable speech rate as the normal speech variant. We segmented speech sounds in the normal and fast variants in each set; we analyzed and compared the duration of vowels and consonants in the two speech rate conditions, as well as the difference, and ratio of long and short vowel pairs in the different conditions using linear mixed effects modeling in R (R Core Team 2018).

Results. On average, word durations in fast speech (199.07±64.69 ms) were half of that found in normal speech (396.76±101.98 ms) with less variability. On segmental durations, we found a SPEECH RATE*SEGMENT TYPE interaction effect ($F(1, 6390) = 103.69$; $p < 0.001$), as in normal speech, vowels were inherently longer and they also reduced more in fast speech than consonants (Fig. 1). On vowel durations, we found a LENGTH*SPEECH RATE interaction effect $F(1, 2095) = 463.34$; $p < 0.001$), since phonologically long vowels were longer in normal speech, and they reduced more in fast speech than short vowels. Additionally, a larger model including V height as a fixed factor also revealed that low vowels, which were inherently longer than high vowels, reduced more in fast speech (LENGTH*HEIGHT*SPEECH RATE interaction $F(1, 2096) = 4.82$; $p < 0.05$) (Fig. 2.). Lastly, in fast speech (compared to normal speech), the differences of all long-short vowel pairs reduced significantly (VOWEL QUALITY*SPEECH RATE interaction: $F(2, 45) = 17.76$; $p < 0.01$),

while the ratio of the long and short vowels' duration decreased only in the /i/-/iː/ pair (Fig. 3. & 4.), as these were differentiated more than the other pairs in normal speech, and ended up being contrasted similar to the other pairs in fast speech (VOWEL QUALITY*SPEECH RATE interaction: $F(2, 75) = 16.86$; $p < 0.01$), while none of the contrasts reduced completely (that is to zero or to 1, respectively; see black dashed lines on Fig. 3. & 4.).

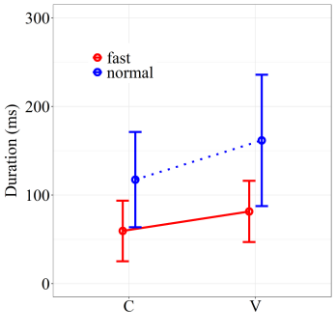


Figure 1: *Vowel and consonant durations in normal and fast speech.*

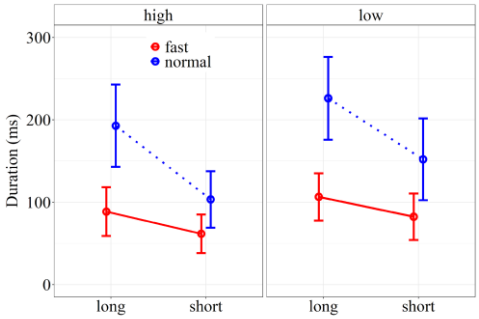


Figure 2: *Vowel durations as a function of phonological vowel length and vowel height.*

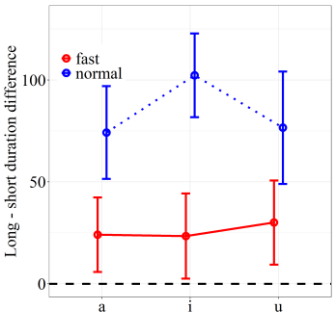


Figure 3: *Differences of long and short vowels' duration as a function of vowel quality.*

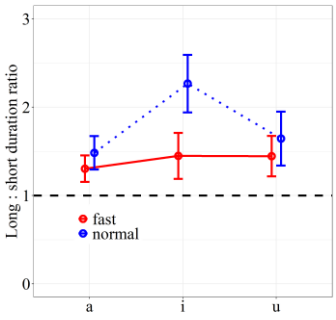


Figure 4: *Ratio of long and short vowels' duration a function of vowel quality.*

Discussion. To conclude, results of our study showed that i) vowels reduced more in their duration than consonants; ii) long vowels showed a greater amount of shortening in fast speech than short vowels, but low vowels, which were inherently longer than high vowels, also showed a greater amount of shortening. Lastly, iii) duration differences of long and short vowels reduced, while duration ratio of the relevant pairs did decrease only in the high front pair. This means that the phonological vowel length contrast was maintained to some extent in fast speech in basically all cases, similarly to Japanese (Hirata 2004). As a next step, we will analyze spectral differentiation of these vowels in the two conditions. These results contribute to our better understanding of how phonological features are implemented in the phonetic realization of speech, and how reduction of segmental features takes place.

Acknowledgements
The research was supported by the TKA–DAAD grant No. 177375., the NKFIH grant No. FK128814, and the ÚNKP-23-3-I-ELTE-335 grant (Zs. Sz.).

References

Boersma, P. & Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3. http://www.praat.org/

Gósy, M. 2004. *Fonetika, a beszéd tudománya.* [Phonetics, the science of speech.] Osiris Kiadó, Budapest.

Greisbach, R. 1992. Reading aloud at maximal speed. *Speech Communication, 11,* 469-473.

Hirata, Y. 2004. Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics, 32,* 565-589.

Kozhevnikov V. A. & Chistovitch L. A. 1965. *Speech articulation and perception.* Joint Publications Research Service, Washington.

Lo R. Y. & Sóskuthy M. 2023. Articulation rate in consonants and vowels: results and methodological challenges from a cross-linguistic corpus study. In Skarnitzl, R. & Volín, J. (Eds.), *Proceedings of the 20th International Congress of Phonetic Science*. Prague: Guarant International. 3206-3210.

Mády, K. 2008. Magyar magánhangzók vizsgálata elektromágneses artikulográffal normál és gyors beszédben. [Analysis of Hungarian vowels using electromagnetic articulography.] *Beszédkutatás*, 52-66.

Magdics K. 1969. A magyar beszédhangok időtartama nyugodt és gyors beszédben. [Duration of speech sounds in Hungarian in calm and fast speech.] *Nyelvtudományi Értekezések, 67,* 45-63.

Magen, H. S. & Blumstein, S. E. 1993. Effects of speaking rate on the vowel length distinction in Korean. *Journal of Phonetics, 21,* 387-409.

R Core Team 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Wood, S. 1973. What happens to vowels and consonants when we speak faster? *Working papers Lund University, Department of Linguistics and Phonetics, 9,* 8-39.

Does the ultrasound probe affect articulatory gestures in children? An acoustic study.

Laura Machart[1,2], Anne Vilain[1], Hélène Lœvenbruck[2] & Lucie Ménard[3]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
[2]Univ. Grenoble Alpes, CNRS, LPNC, 38000 Grenoble, France
[3]Laboratoire de Phonétique, UQÀM, Montréal, Canada

laura.machart@univ-grenoble-alpes.fr, anne.vilain@univ-grenoble-alpes.fr,
helene.loevenbruck@univ-grenoble-alpes.fr, menard.lucie@uqam.ca

Introduction. Ultrasound imaging of the sagittal profile of the tongue provides objective information on the degree of articulatory precision (Ménard et al., 2014). Although studying articulatory gestures by using ultrasound is not an invasive technique (Ménard et al., 2012; Turgeon et al., 2017), placing the probe under the child's chin can disturb the production of a number of phonemes, in particular by slightly constraining jaw movements. A study by Villegas et al. (2015), found a small but not significant effect of the ultrasound probe on jaw movements, with over-articulation with the ultrasound probe. Pucher et al. (2020), for their part, have shown that some stabilization headsets may influence formant values in vowel production. Very few data are currently available on the potential effect of the ultrasound probe on speech production in adults. Even fewer data are available for children, although Machart et al. (accepted) have shown that the use of ultrasound to study articulatory movements in this population is an informative method. As part of this previous speech production study in children with typical hearing and with cochlear implants, a comparison between the acoustic nature of sounds produced without the ultrasound probe (noUS modality) and with the ultrasound probe (US modality) was performed to ensure the reliability of the results in the ultrasound modality. The present paper, therefore, aims to discuss the potential impact of the ultrasound probe on articulatory gestures in children.

Methods. The lingual movements of ten children with typical hearing (TH group) and eight children with cochlear implants (CI group) were recorded during a picture-naming task. All of the children were native speakers of Canadian French. At the time of the experiment, children in the TH group were aged 52 to 137 months (mean age = 96.25 months, sd = 25.68) and children in the CI group were aged 65 to 133 months (mean age = 102.55, sd = 19.35). The corpus consisted of four words, each including one of the four targeted consonants /t/, /k/, /s/ or /ʃ/. The chosen words were disyllabic and included the targeted consonant in initial position and followed by vowel /a/, to ensure a higher articulatory precision. This resulted in the four French words: 'tapis' /tapi/ *carpet*, 'carotte' /kaʁɔt/ *carrot*, 'sapin' /sapɛ̃/ *fir tree* and 'chapeau' /ʃapo/ *hat*. Stimuli were chosen for their lexical frequency and imageability. Each word was produced six times in the carrier sentence 'C'est les…' /sele/ '*These are…'* and prompted with six different pictures. A total of 24 tokens per condition was recorded (six repetitions of the four target words), minimizing fatigue while ensuring a sufficient number of repetitions. The order of the stimuli was randomized between participants and across conditions. The first condition (noUS modality) consisted in recording the child's speech production without the ultrasound probe. In the second condition (US modality), the ultrasound probe was added under the child's chin and the 24 tokens were recorded again, in order to provide acoustic and articulatory data. All the acoustic data (i.e., 48 items per child) were then phonetically transcribed using PRAAT (Boersma & Weenink, 2019). The formant values of F2 and F3 at consonant offset were extracted for stop and fricative consonants. The first spectral moment (i.e., mean center of gravity) was also extracted for fricative consonants. Comparisons between the acoustic measurements in the US and noUS modalities were then run for each group (i.e., TH and CI groups) using linear mixed effects models (*lme* function in R, R Core Team, 2023).

Results. As concerns stop consonants, a trend effect of the ultrasound probe is observed on F2 values in both groups ($p=.074$ in the TH group and $p=.077$ in the CI group). This trend effect is also observed on F3 values in the CI group ($p=.098$). Further investigations reveal that the limitation of jaw movements by the ultrasound probe significantly affects the production of the F2 values of five TH participants whereas only one CI participant seems to be perturbated (Figure 1a). With regard to fricative consonants, a significant effect of the ultrasound probe is observed on F2 values in both groups ($p<.001$ in the TH group and $p=.012$ in the CI group). Adding the ultrasound probe significantly modifies the F2 values of six TH participants and three CI participants (Figure 1b).
However, all these differences observed between modalities do not seem to influence the distinction between places of articulation: formant values tend to be lower with the addition of the ultrasound probe but the difference between places of articulation of stop and fricative consonants is not altered ($p<.001$ in both modalities for stop and fricative consonants). This reduction in F2 can be interpreted as a tongue withdrawal to modulate the limited movement of the jaw.

Discussion. These results show that the limitation of jaw movements induced by an ultrasound probe does not seem to interfere with the distinction between places of articulation in stop and fricative consonants. However, they also confirm that adding an ultrasound probe under a child's chin may alter speech production. Moreover, the ultrasound probe not only perturbates speech production of children with atypical development but also children with typical development, and the amplitude of perturbation varies significantly from one individual to another. This study, therefore, pleads for the use of acoustic measurements to complement ultrasound imaging of the tongue. We suggest to systematically record a condition without the ultrasound probe to measure its impact on speech production. This may highlight participants who are more sensitive to probe disturbance, and whose modified productions could alter the relevance of articulatory measurements.

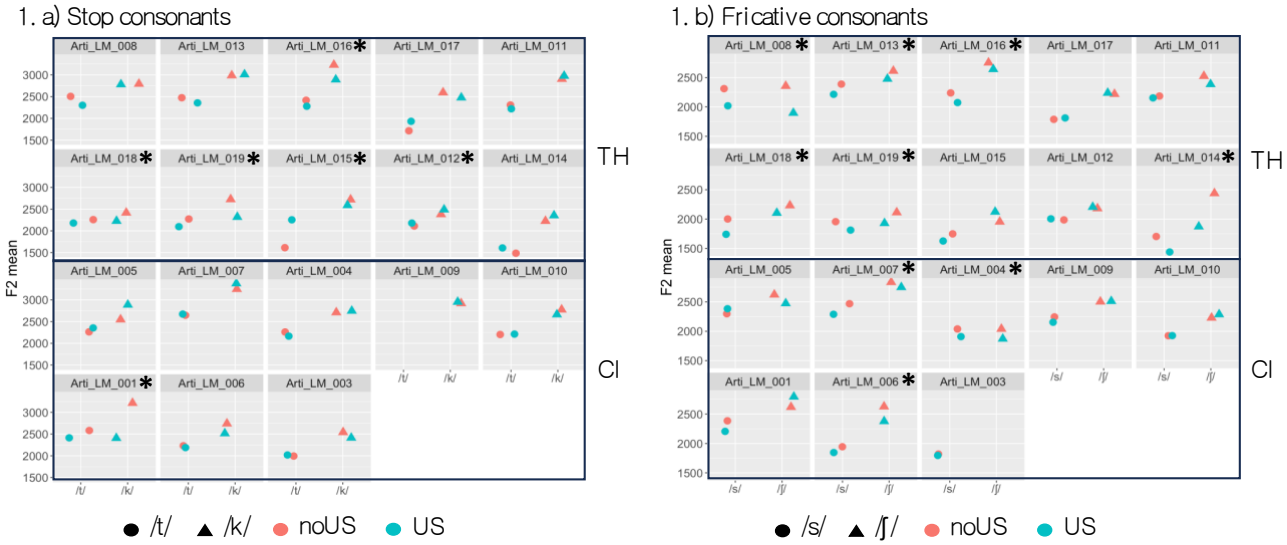


Figure 1: *Differences between modalities for F2 values. Participants are listed by group (TH in the two upper lines) and ordered by chronological age (youngest to oldest). Black stars indicate a significant difference between the US (blue) and noUS (red) modalities.*

References

Boersma, P., & Weenink, D. (2019). *Praat: Doing Phonetics by Computer*. https://www.fon.hum.uva.nl/praat/

Machart, L., Vilain, A., Lœvenbruck, H., Tiede, M. & Ménard, L. (accepted). Exposure to Canadian French Cued Speech improves consonant articulation in children with cochlear implants: acoustic and articulatory data. *Journal of Speech, Language and Hearing Research*, special issue from the 8th International Conference on Speech Motor Control, Groningen, Netherlands.

Ménard, L., Aubin, J., Thibeault, M., & Richard, G. (2012). Measuring Tongue Shapes and Positions with Ultrasound Imaging: A Validation Experiment Using an Articulatory Model. *Folia Phoniatrica et Logopaedica*, *64*(2), 64–72.

Ménard, L., Leclerc, A., & Tiede, M. (2014). Articulatory and acoustic correlates of contrastive focus in congenitally blind adults and sighted adults. *Journal of Speech, Language, and Hearing Research*, *57*(3), 793–804.

Pucher, M., Klingler, N., Luttenberger, J., & Spreafico, L. (2020). Accuracy, recording interference, and articulatory quality of headsets for ultrasound recordings. *Speech Communication*, *123*, 83–97.

R Development Core Team. (2023). *RStudio | Open source & professional software for data science teams*. https://www.rstudio.com/

Turgeon, C., Trudeau-Fisette, P., Fitzpatrick, E., & Ménard, L. (2017). Vowel intelligibility in children with cochlear implants: An acoustic and articulatory study. International *Journal of Pediatric Otorhinolaryngology*, 101, 87-96.

Villegas, J., Wilson, I., Iguro, Y., & Erickson, D. (2015). Effect of a fixed ultrasound probe on jaw movement during speech. *Proc. Ultrafest VII.*

SPRAAKLAB: mobile laboratory to collect high-quality speech data

Martijn Wieling[1], Teja Rebernik[1], Jidde Jacobi[1], Thomas Tienkamp[1], Frank Tsiwah[1], Defne Abur[1]

[1]University of Groningen, Faculty of Arts

{m.b.wieling, t.rebernik, j.jacobi, t.b.tienkamp, f.tsiwah, d.abur}@rug.nl

Introduction. In speech production and perception research, data collection is often conducted in laboratory rooms of a research institute. This may pose a high participation threshold for some participants and may thereby introduce a selection bias. Unsurprisingly, most linguistic studies therefore tend to include "a homogeneous population of compliant undergraduates" (Whalen & McDonough, 2015, p. 397). Even when studies do not target university students, they still tend to test within the walls of the research institute. Of course, this may not be optimal if the goal is to study a more dispersed population, or if the study is targeting a specific group that is not likely to frequent the university (e.g., older adults, or individuals with a speech disorder). One solution to lowering the threshold for participation is to use portable acoustic and articulatory recording equipment (Whalen and McDonough, 2015) and visit speakers at their home (e.g., Wieling et al., 2016). Unfortunately, this approach may negatively impact the quality of the collected data, as the location's characteristics may affect the quality of acoustic (due to background noise) and (electromagnetic) articulatory-kinematic recordings (due to metal objects in the vicinity). Instead, in this abstract we propose to bring a full laboratory environment to the speaker. In the following paragraphs, we will discuss some of the specifications of our mobile laboratory, SPRAAKLAB, and also show that data collected in the mobile laboratory is not of lower quality than data collected in a regular university laboratory.

Mobile laboratory specifications. SPRAAKLAB (Speech Recorded Acoustically And Kinematically LABoratory) is a large van (L x W x H: 7m x 2.75m x 3m) featuring an attractive outside design reflecting the various research techniques we use. Through generous funding (€150,000) of our university, it was custom-built on top of a lowered chassis of a Fiat Ducato at the end of 2020.[1] As the weight of the van is below 3,500 kg, it can be driven with a regular EU B-class driver's license. As it runs of Diesel fuel, its range is substantial and hence can be used for any target populations accessible by road from the Netherlands (and where are gas stations within a range of about 600 km). The mobile laboratory contains two separate rooms: one where the experimenter(s) can control all equipment, and a sound-dampened (-40dB dampening) room in which the acoustic and articulatory recordings can be made. Figure 1 shows the outside and inside of SPRAAKLAB. Sound dampening has been achieved by using bitumen and wood in the construction, instead of metal. As a consequence, the environment does not interfere with the functionality of an (NDI-VOX) electromagnetic articulography (EMA) system (see Rebernik et al., 2021).

Figure 1: *SPRAAKLAB mobile laboratory. Top-left and top-right: outside. Bottom-left and bottom-middle: experimenter room. Bottom-right: sound-dampened room.*

[1] The running costs consisting of fuel, etc. of approx. €15,000 per year, are covered by the Faculty of Arts of our university.

Mobile laboratory use. The setup of the SPRAAKLAB is highly flexible. It allows both for the collection of speech articulation and acoustic data of individual speakers, as well as for acoustic speech data collection in dyads. In addition to using the mobile laboratory to collect data at participant's homes, we also use it to be present at music and science events. At these events, we inform the general audience about the type of research we conduct, and how we do this. For example, we show them the movement of their tongue during speech with our ultrasound tongue imaging (UTI) system. Besides demonstrating these aspects, we also collect research data in short (< 20 minute) experiments. Specifically, in this abstract, we will compare data collected in a 10-minute formant-perturbation experiment at the two-week theatre festival *Noorderzon* to data collected using a comparable setup in our regular university laboratory.

Methods. A total of 41 participants (24F, 17M, age: 22-59 years) participated in this study, of which 20 speakers (mean age 31.1 years, 10M, 10F) participated at the *Noorderzon* festival and 21 (mean age 36.3 years, 6M, 14F) in a traditional laboratory at the University of Groningen. The participants completed a gradual formant perturbation experiment, in which they pronounced six different target words with the open-mid front unrounded vowel /ɛ/ for 114 trials following a common setup for adaptation experiments (e.g., Cai et al., 2010; 24 start trials, 24 ramp trials, 48 stay trials and 18 end trials) in which the vowel was shifted upwards (towards /ɪ/) by decreasing the first formant (F_1) by 20% and increasing the second formant (F_2) by 15%. After the experiment, we calculated mean F_1 and F_2 across 30-80 ms of the vowel trajectory for each vowel production to capture the stable mid-point of the vowel. We used generalized additive mixed-effects regression modeling (see Wieling, 2018) to compare the patterns over trial (we did not aggregate per phase) between the two different laboratory settings. Before analysis, all formant measures were *z*-transformed per individual based on the individual's mean and standard deviation values during the START phase.

Results. Figure 2 visualizes the results per formant. For F_1, the difference between laboratories was not significant (p = 0.7). For F_2, the SPRAAKLAB participants showed a significantly (α: 0.05) greater compensation (p = 0.04).

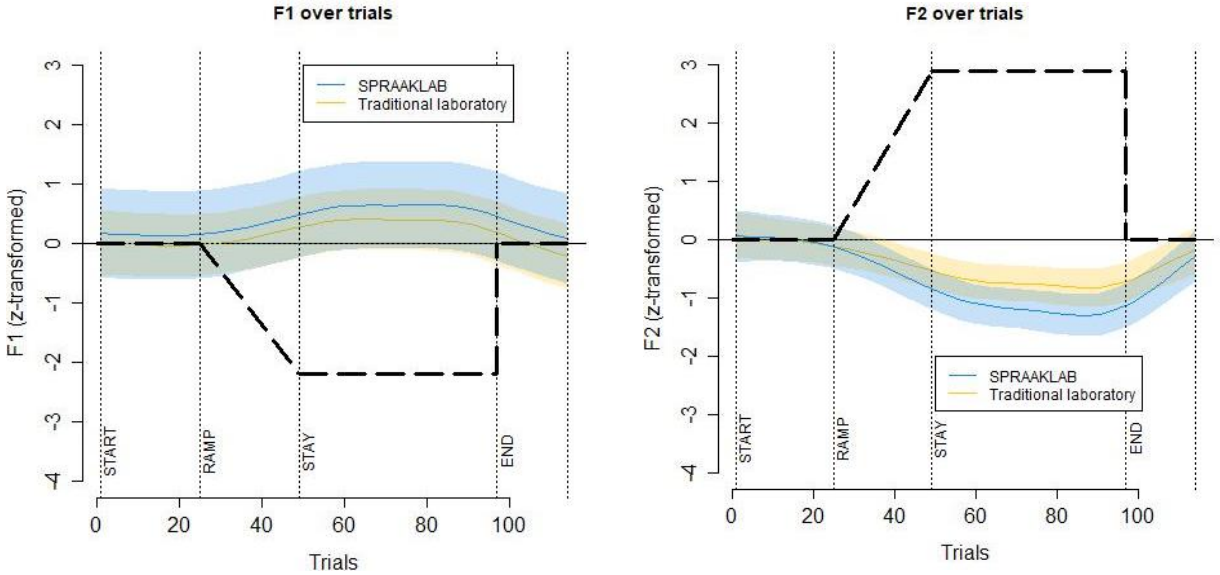


Figure 2: Results of statistical analysis using generalized additive mixed-effects regression modeling. Dotted lines indicate the separation into the four phases. The dashed line indicates the average normalized perturbation applied to participants' vowel productions.

Discussion. The results of our study comparing the two laboratory settings suggest that test results in SPRAAKLAB are comparable to test results in a traditional laboratory. In sum, SPRAAKLAB appears to be suitable as a laboratory when running (acoustic) experiments requiring a strict experimental control.

References

Cai, S., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2010). Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong/iau/and its pattern of generalization. *The Journal of the Acoustical Society of America*, 128(4), 2033-2048.

Rebernik, T., Jacobi, J., Tiede, M., & Wieling, M. (2021). Accuracy assessment of two electromagnetic articulographs: Northern digital inc. wave and northern digital inc. vox. *Journal of Speech, Language, and Hearing Research*, 64(7), 2637-2667

Whalen, D. H., & McDonough, J. (2015). Taking the laboratory into the field. *Annu. Rev. Linguist.*, 1(1), 395-415.

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S.N., & Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics*, 59, 122-143.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, 70, 86-11.

Presenting ADA: A Tool for Articulatory Data Analysis

Philipp Buech, Anne Hermes

Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle)

philipp.buech@sorbonne-nouvelle.fr, anne.hermes@sorbonne-nouvelle.fr

Introduction. Speech production, more specifically articulation, can be investigated by a variety of techniques. Electromagnetic articulography (EMA) is a minimally invasive technique that is widely used in the field (Rebernik et al., 2021). Kinematic movements can be recorded by tracking the position of sensors attached to e.g., articulators. Researchers have multiple tools at their disposal for the display and annotation of EMA data, e.g., MView (Tiede, 2005), VisArtico (Ouni et al., 2012), emuR (Winkelmann et al., 2007), or EMATOOLS (Nguyen, 2000). Some tools may no longer be maintained and certain prerequisites may be necessary for their use, such as external licenses (e.g., MATLAB) or a certain level of programming skills which may be a barrier for non-technical users. Recently, further tools were developed to make EMA data available in Praat (Boersma & Weenink, 2023), e.g., ema2wav (Buech et al., 2022) and Kijk (Machado & He, 2023), which allow the display and manual annotation of articulatory trajectories. However, their use may require certain tweaks in Praat such as muting the channels which contain the articulatory data and visualizing only one dimension at a time per channel, thus the visualization and annotation of multiple dimensions may be overwhelming. Here, we present ADA, a tool for articulatory data analysis, applicable for post-processing, display, (automatic) annotation and measurement of kinematic data collected via EMA. ADA is a free and open-source software that is implemented in Python (Rossum & Drake, 2009) and thus runs on Mac, Windows, and Linux. It offers a comprehensive, user-friendly Graphical User Interface (GUI) and with all its features presents a remarkable advancement for working with articulatory data.

Features. The current version of ADA allows the input of (i) EMA data, (ii) corresponding audio recordings and (iii) annotation files. (i) EMA data can be either loaded from the .pos files of the AG200 and AG500/501 models of Carstens Medizinelektronik GmbH or as external data (in .csv or .tsv format). (ii) Audio is supported for .wav, .mp3 and .ogg files and (iii) annotations can be loaded from .TextGrid, .json or .lab files. Once the data is loaded, users can optionally choose between a moving average or a Butterworth filter for the articulatory signal. The view of the sensor trajectories over time is shown in Fig. 1 (left). The waveform of the corresponding audio recording and a specific tier of the audio annotations, if available, are displayed at the top of the window. The kinematic data can be displayed simultaneously on three different panels, each with three possible axes. Plotting options include not only the display of the sensor trajectories, but also their derivatives like velocity, acceleration, and tangential velocity, as well as the sensor distances in one dimension or the two-dimensional and three-dimensional Euclidean distances. Regarding annotation, landmarks can be added either manually or automatically. The manual annotation covers adding, adjusting and/or deleting landmarks in their position or label. To determine landmarks automatically, a region of interest (ROI) can be selected on each panel. Here, ADA provides multiple approaches for automatic landmark detection: these are based on the velocity or tangential velocity with a threshold of 20% (Kroos et al., 1997) or 15% (Chitoran et al., 2002) of the peak velocity or based on the acceleration profile. Per default, the gesture's onset, onset speed maximum, target achievement, release, offset speed maximum and offset are detected. Besides the display of the trajectories, users can view the sensor positions in a two-dimensional or three-dimensional space. Fig. 1 (right) shows an example of the two-dimensional view. In this viewing mode, users can choose which dimensions to display (horizontal-vertical, vertical-lateral, horizontal-lateral). In the figure, the horizontal dimension is plotted on the x-axis and the vertical dimension on the y-axis, along with the positions of the LLIP, ULIP, TTIP, TMID and TBO sensors. By hovering with the cursor over the acoustic waveform the sensor positions are changing/moving on-line. Furthermore, the tongue shape can be visualized by as cubic spline interpolation. If a region of interest in the corresponding waveform is selected, the sensor movements can be displayed (see Fig. 1, right). Importantly, the automatic detection of landmarks can be done across files (if acoustics segmentation is available), thus, it is not necessary to view each file separately. Users can determine a set of segments or sequences of segments that will be annotated automatically across all uploaded files. Further, these landmarks can be inspected and modified as described above. Likewise, there is also the option available to extract specific kinematic parameters (e.g., duration, displacement, stiffness) and sensor trajectories. Finally, ADA allows to export EMA data and landmark annotations in various formats. EMA data can be exported as .csv files or as self-describing data sets in the Network Common Data form (NetCDF) format. The articulatory landmarks can be exported into .csv or .lab files, or as point tiers in .TextGrid files. Besides the

user-friendly GUI-based approach, technical users can use all functions within ADA for data import, filtering, landmark detection, measurements, and also data export in their own custom Python scripts.

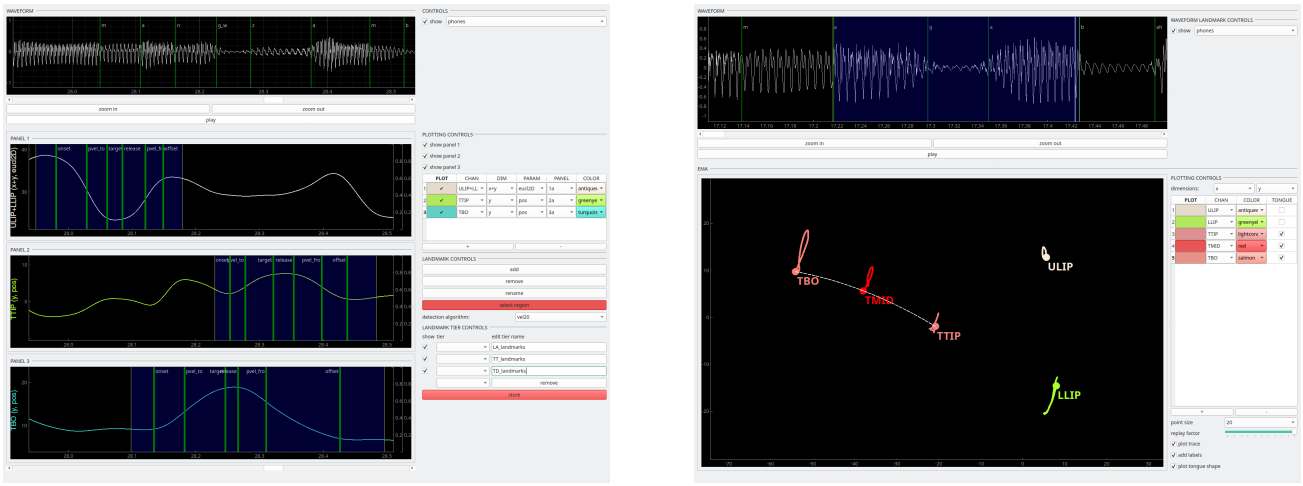


Figure 1: *Left: ADA's display of waveform and three trajectories (Euclidean distance of ULIP and LLIP, and vertical position of TTIP and TBO sensors), ROIs and automatically detected landmarks. Right: ADA's two-dimensional view of a ROI in an [aga] sequence with waveform (top), sensor positions, labels, interpolated tongue shape and movement track (bottom).*

Conclusion. ADA's key strength lies in its accessibility and versatility. It is a free, open-source tool that operates on multiple platforms and requires no programming skills. This makes it an invaluable resource for student, researchers, technical and non-technical users alike. The software supports a wide range of data formats and it also provides multiple options for data visualization (sensor trajectories and positions in the two-dimensional and three-dimensional space) and annotation, including manual alignment and automatic landmark detection for individual gestures and sets of gestures across files and with multiple approaches. Further, ADA allows for the extraction of kinematic parameters and movements, as well as the export of EMA data and landmark annotations in various formats. In conclusion, ADA is a powerful tool that simplifies the analysis of EMA data, making it more accessible and efficient for everyone within this research area. It is planned to add also support for the AG100 and NDI WAVE/VOX systems, as well as to integrate a head-correction procedure.

References.

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer* [Computer program] (6.4.01). Retrieved November 30, 2023, from http://www.praat.org/

Buech, P., Roessig, S., Pagel, L., Muecke, D., & Hermes, A. (2022). Ema2wav: Doing articulation by Praat. *Proc. Interspeech 2022*, 1352–1356. https://doi.org/10.21437/Interspeech.2022-10813

Chitoran, I., Goldstein, L., & Byrd, D. (2002). Gestural overlap and recoverability: Articulatory evidence from Georgian. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (pp. 419–448). De Gruyter Mouton. https://doi.org/10.1515/9783110197105.2.419

Kroos, C., Hoole, P., Kühnert, B., & Tillmann, H. G. (1997). Phonetic evidence for the phonological status of the tense-lax distinction in German. In *Fipkm* (pp. 17–25, Vol. 35).

Machado, C. L., & He, L. (2023). Kijk: A praat plugin to visualise articulatory trajectories. In R. Skarnitzl & J. Volín (Eds.), *Roceedings of the 20th international congress of phonetic sciences* (pp. 4115–4119). Guarant International.

Nguyen, N. (2000). A MATLAB toolbox for the analysis of articulatory data in the production of speech. *Behavior Research Methods, Instruments, & Computers*, *32*(3), 464–467. https://doi.org/10.3758/BF03200817

Ouni, S., Mangeonjean, L., & Steiner, I. (2012). Visartico: a visualization tool for articulatory data. *Proc. Interspeech 2012*, 1878–1881. https://doi.org/10.21437/Interspeech.2012-510

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology*, *12*(1), 1–42. https://doi.org/10.5334/labphon.237

Rossum, G. V., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Tiede, M. (2005). *MView: Software for the visualization and analysis of concurrently recorded movement data*. Haskins Laboratory.

Winkelmann, R., Harrington, J., & Jänsch, K. (2007). EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, *45*, 392–410.

On The Utility of a Single-Breath Counting Task for the Remote Digital Assessment of Respiratory Function in ALS

Michael Neumann[1], Hardik Kothare[1], and Vikram Ramanarayanan[1,2]

[1]*Modality.AI, Inc.*
[2]*University of California, San Francisco*
michael.neumann@modality.ai

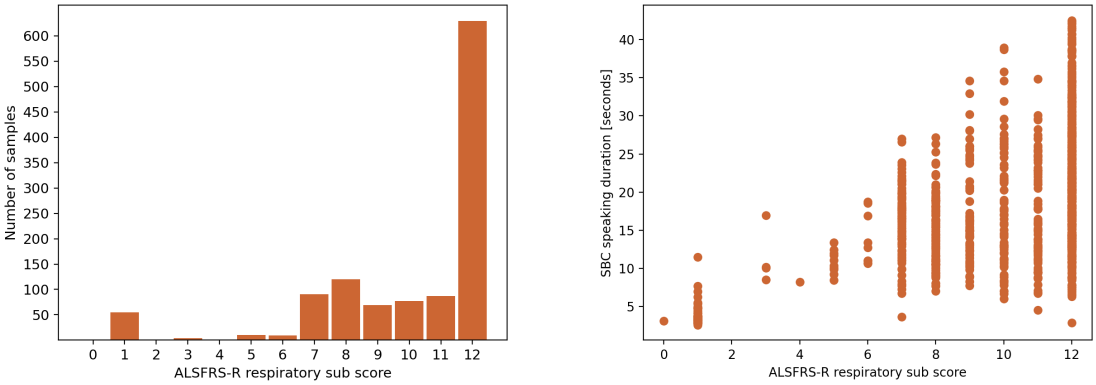
Introduction and Objectives. Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disease, characterized by degeneration of lower and upper motor neurons. Early detection of respiratory decline is crucial for optimal symptom management for people with ALS (Hardiman 2011). Respiratory function is commonly evaluated with a spirometer, both in clinic and remote (Baroi et al. 2018). However, such tests need the help of a caregiver in most cases (Tattersall et al. 2022) and can be difficult to execute for people with faciobulbar weakness (Lechtzin et al. 2018).
The demand for low-burden objective digital biomarkers is high, both to enhance clinical trials and to improve patient monitoring. It has been shown that forced vital capacity can be estimated from speech acoustics, specifically from sustained vowel phonation (Stegmann et al. 2021; Tabor Gray et al. 2023). In this work, we explore single breath counting (SBC) as an alternative assessment of respiratory function. SBC is a more ecologically valid assessment of daily function compared to the artificial nature of the sustained vowel phonation test, because it involves natural speech articulation. Previous research on SBC in clinical settings has shown that the duration correlates with standard pulmonary and respiratory measurements (Bartfield et al. 1994; Ali et al. 2011). In this work, we assess the feasibility of transferring this task into a self-driven remote assessment, which is based on a web based dialog system.

Methods. A web based dialog system (Ramanarayanan et al. 2023) was used to collect speech recordings from participants. For the SBC task, participants were instructed to take a deep breath and count up from one until they run out of breath. Additionally, participants filled out the ALS functional rating scale - revised (ALSFRS-R) (Cedarbaum et al. 1999). The ALSFRS-R consists of 12 questions that capture functional impairment in four domains. Each question can have a score between 0 and 4, where 4 indicates full function. We used the respiratory sub score (three questions about dyspnea, orthopnea, and respiratory insufficiency; range from 0 to 12) to investigate the correlation between SBC duration and respiratory function. The SBC duration was computed automatically for every sample using Praat (Boersma 2001). We computed two metrics: *speaking duration* (including silences within the utterance), and *articulation duration* (excluding silences, i.e. the duration of all speech events concatenated). We have shown previously that robust automatic articulation boundary detection is feasible in this remote setting where participants use their own devices (Liscombe et al. 2022). We report Spearman correlation between SBC duration and the ALSFRS-R respiratory sub score. Additionally, a non-parametric Kruskal-Wallis test was done on a cross-sectional subset of the data to test whether SBC duration differs statistically between participants with a respiratory score of 12 (*RES_12*) and participants with a score below 12 (*RES<12*). For this, every participant's first sample was considered.

Data and Demographics. Recordings from 96 people with ALS (46 females, mean age (SD): 62.1 (8.9) years) were collected between 2021-12-02 and 2024-02-01 in collaboration with EverythingALS and the Peter Cohen Foundation[1]. The study protocol was granted exempt status by an external Institutional Review Board[2]. The total number of sessions in the dataset is 1,153 (54.6% with a respiratory sub score of 12, see Fig. 1a).

Results. The Spearman correlation coefficient between respiratory sub score and SBC duration was 0.43 ($p < 0.0001$) for *speaking duration* and 0.44 ($p < 0.0001$) for *articulation duration* when considering all samples, and 0.37 ($p < 0.0001$) for *speaking duration* and 0.38 ($p < 0.0001$) for *articulation duration* for the *RES<12* group. Figure 1b shows the relationship between SBC duration and the respiratory sub score. SBC *articulation duration* was significantly different

[1]https://www.everythingals.org/research
[2]https://www.advarra.com/



(a) (b)

Figure 1: (a) Distribution of ALSFRS-R respiratory sub score. The *RES<12* group has a median score of 8.0 with a standard deviation of 2.8. (b) Relationship between SBC duration and respiratory sub score.

between the two groups *RES_12* and *RES<12* at $p < 0.05$. The effect size in terms of Glass' delta was moderate (-0.46), indicating a shorter mean duration in the *RES<12* cohort.

Discussion. We examined the feasibility of administering the SBC task within a web based remote speech assessment and its utility to capture information on respiratory function. We have shown that the SBC duration has a moderate correlation with the self-reported ALSFRS-R respiratory sub score. The correlation is higher at the lower end of the distribution (score of 6 and below), whereas we observed large variation in SBC duration in samples with higher respiratory scores. An important observation during the study was that participants performed the task in different ways; some counted at a fast pace, while others counted slow with distinct pauses between numbers. This emphasizes the importance of clear and unambiguous instructions in such a self-driven assessment. Another caveat is the use of the self reported ALSFRS-R respiratory sub score as a proxy for respiratory function. Future work should involve spirometry measurements as ground-truth. Lastly, we acknowledge the fact that pure correlation with the ALSFRS-R respiratory sub score can be misleading in individual cases, as we observed participants with a constant respiratory sub score for whom the SBC duration decreased over time. This suggests the possibility of detecting changes early, before they are reflected in the ALSFRS-R responses.

References.

Ali, Syed Sameer et al. (2011). "Single-breath counting: a pilot study of a novel technique for measuring pulmonary function in children". In: *The American journal of emergency medicine* 29.1, pp. 33–36.

Baroi, Sidney et al. (2018). "Advances in remote respiratory assessments for people with chronic obstructive pulmonary disease: a systematic review". In: *Telemedicine and e-Health* 24.6, pp. 415–424.

Bartfield, Joel M et al. (1994). "Single breath counting in the assessment of pulmonary function". In: *Annals of emergency medicine* 24.2, pp. 256–259.

Boersma, Paul (2001). "Praat, a System for Doing Phonetics by Computer". In: *Glot International* 5.9/10, pp. 341–345.

Cedarbaum, Jesse M et al. (1999). "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function". In: *Journal of the neurological sciences* 169.1-2, pp. 13–21.

Hardiman, Orla (2011). "Management of respiratory symptoms in ALS". In: *Journal of neurology* 258.3, pp. 359–365.

Lechtzin, Noah et al. (2018). "Respiratory measures in amyotrophic lateral sclerosis". In: *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 19.5-6, pp. 321–330.

Liscombe, Jackson et al. (2022). "On the Robust Automatic Computation of Speaking and Articulation Duration in ALS Patients Versus Healthy Controls". In: *Proceedings of the Motor Speech Conference, Charleston, SC*.

Ramanarayanan, Vikram et al. (2023). "When Words Speak Just as Loudly as Actions: Virtual Agent Based Remote Health Assessment Integrating What Patients Say with What They Do". In: *Proc. INTERSPEECH 2023*, pp. 678–679.

Stegmann, Gabriela M et al. (2021). "Estimation of forced vital capacity using speech acoustics in patients with ALS". In: *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 22.sup1, pp. 14–21.

Tabor Gray, Lauren et al. (2023). "Maximum Phonation Time as a Surrogate Marker for Airway Clearance Physiologic Capacity and Pulmonary Function in Individuals With Amyotrophic Lateral Sclerosis". In: *Journal of Speech, Language, and Hearing Research* 66.4, pp. 1165–1172.

Tattersall, Rachel et al. (2022). "The patient's perspective of remote respiratory assessments during the COVID-19 pandemic". In: *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 23.1-2, pp. 76–80.

Auditory-motor adaptation of vowels across adulthood

Katharina Polsterer[1], Thomas Tienkamp[1], Teja Rebernik[1], Hedwig Sekeres[1], Valentine Lucquiault[1], Martijn Wieling[1], Defne Abur[1]

[1]Centre for Language and Cognition Groningen, University of Groningen

k.m.polsterer@rug.nl

Introduction. Aging is associated with a number of changes relevant to speech production including cognitive (MacPherson 2019) and sensory decline (Jones & Noppeney 2021). Further, speech production itself changes with age, as movements become more asymmetrical in timing, slower, and more variable (Hermes et al. 2018; Tremblay et al. 2017). Moreover, neuroimaging shows spatial differences in cortical activations during speech production of younger (mean age 26.8 years) and older adults (mean age 68.2 years; Tremblay et al. 2017). The brain constantly monitors the alignment of speech motor commands and consequent auditory feedback in order to maintain accurate and fluent speech (Guenther 2016). This control mechanism enables the speaker to gradually adapt motor commands in response to mismatches of the expected and the received auditory feedback (Houde & Jordan 1998), which is referred to as auditory-motor adaptation. The sum of motor, cognitive, and sensory decline with advancing age suggests that auditory-motor adaptation may similarly undergo a decline. Indeed, aging has been found to affect auditory-motor control of pitch, such that the magnitude of reflex-like vocal-motor compensation in response to auditory feedback perturbation decreases after the fifth decade of life (Liu et al. 2011). However, it is unclear whether aging affects articulatory auditory-motor control as well. The current study therefore investigated adaptive responses to auditory perturbations of vowel articulation across age in typical adulthood. We hypothesized that aging is related to decreasing auditory-motor adaptation as closely related functions decline. Specifically, we predicted that adaptive responses to perturbations of the first formant (F_1) decrease in magnitude with older compared to younger speakers.

Methods. A total of 81 first-language speakers of Italian (46 women, 35 men) participated in this study. The participants were between 16 and 82 years old (mean age 39 years ± 16.9 SD). They indicated no speech, language, or neurological disorders, and passed a pure-tone hearing screening (< 50 years of age: 25 dB HL at 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz; > 50 years of age: 25dB HL at 250 Hz, 500 Hz, 1 kHz, and 40 dB HL at 2 kHz, 4 kHz). The participants were recorded in a sound attenuated booth inside a research van (Wieling et al. 2023). They were wearing headphones (Sennheiser HD 280 Pro) and a microphone (Shure MX153), which was placed at a 7 cm distance from the mouth.
For the experimental task, the participants repeatedly produced the three Italian words /ˈbeːve/ (‘drink’), /ˈdeːve/ (‘must’), and /ˈveːde/ (‘see’) in random order, when prompted by text on a screen. While speaking, the participants received real-time auditory feedback via the headphones, which was amplified by 5 dB. The experiment comprised a total of 108 trials split over four phases: baseline, ramp, hold, and after-effect. During the 24 baseline trials, the participants received unperturbed auditory feedback. Over the course of the 30 subsequent ramp trials, the first formant (F_1) in the auditory feedback was gradually in equal steps until it reached 50% increase relative to the baseline mean F_1. For the 30 trials of the hold phase, the F_1 increase was held constant at 50% relative to the mean F_1 of the baseline phase. During the last 24 trials, constituting the after-effect phase, there was no perturbation. Formant shifting was done using Audapter (v2.1.012; Cai et al. 2008). In post-processing, the first vowel of each trial was selected manually on the spectrogram, and the mean F_1 in Hz was measured in a window of 40 to 120 ms of that selection. F_1 values were then normalized per trial as percentage change (% change) relative to the baseline mean F_1. Response magnitudes were calculated as mean % change of F_1 across the hold phase.

Results. An initial inspection of our data indicated that individually, participants either showed negative change in F_1 (i.e., ‘opposing’ the perturbation), positive change in F_1 (i.e., ‘following’ the perturbation), or no change in F_1 (i.e., ‘non-responding’ to the perturbation). Therefore, we split our data into response types. Through two-tailed one-sample *t*-tests, we determined for each individual participant, whether % changes of F_1 in the trials of the hold phase (30 trials per participant) were significantly different from 0, given an alpha level of 0.05. Participants with a non-significant result were categorized as ‘non-responding’ (n = 17). Results that were significantly lower than 0 were classified as ‘opposing’ (n = 31), while results that were significantly greater than 0 were classified as ‘following’ (n = 33). The three groups of participants, each associated with a different response type, did not significantly differ in age, as indicated by linear modelling (opposing: β = -0.213, p = 0.114; following: β = 0.151, p = 0.254; non-responding as the reference level). To address our hypothesis, we then tested the (potentially non-linear) effect of age on response magnitudes using generalized additive modelling (cf. Wieling 2018). Model comparison showed that a model including an interaction of age and

response types as predictor was to be preferred over a model with age across response types (without the interaction; $p < 0.001$). This indicated different aging patterns per response type. The results of the preferred model, including the interaction, showed a significant effect of age on response magnitude (opposers: $F = 6.010$, $p = 0.017$; followers: $F = 6.512$, $p < 0.001$; non-responders as the reference level), as presented in **Figure 1**.

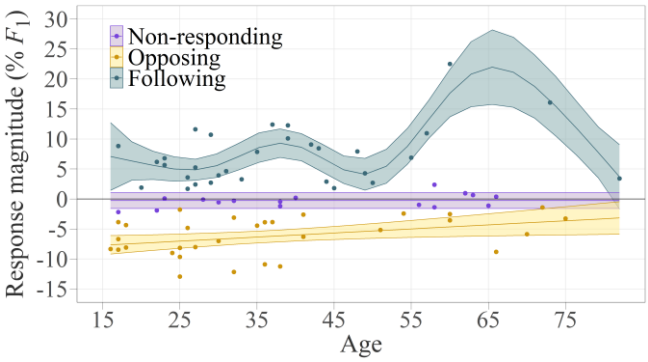


Figure 1: *Mean % F_1 change in response to 50% F_1 increase in the auditory feedback during the hold phase.*

Discussion. This study aimed to shed light on auditory-motor adaptation of vowel articulation across adulthood in typical aging. Not every participant opposed the auditory perturbation in their change of F_1, as many followed the perturbation. Others did not seem to adapt their articulation in response to the auditory perturbation at all. Previous research has described similar differences in response types, although in the pitch perturbation domain (Behroozmand et al. 2012; Franken et al. 2018). While this division is puzzling, age does not seem to be a driving factor since we found all response types to occur across age. Thus, other factors besides aging are probably at play triggering different response types. However, the results suggest that with advancing age, response magnitudes change separately for different response types. Opposing responses tend to gradually become smaller in magnitude, as was initially hypothesized, potentially indicating reduced auditory-motor integration. Following responses also seem to vary by age, which appears to be driven by only few data points at older age. In the group of non-responding participants, no effect of age was observed. However, since data of older participants is scarce, especially compared to younger participants, additional data is needed to further clarify the aging pattern, particularly for following responses. Previous results on auditory-motor control of pitch suggest a systematic decrease in magnitudes of opposing responses after the fifth decade of life (Liu et al. 2011). We extend these findings by showing an age-related decrease in magnitudes of opposing responses in the articulatory domain. These were however not found to occur in a specific age range, but showed in a gradual manner across age.

References

Behroozmand, R., Korzyukov, O., Sattler, L., & Larson, C. R. (2012). Opposing and following vocal responses to pitch-shifted auditory feedback: evidence for different mechanisms of voice pitch control. *J Acoust Soc Am*, *132*(4), 2468–2477.

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2008). *A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/* [Paper presentation]. 8th International Seminar on Speech Production, Strasbourg, France.

Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., & Eisner, F. (2018). Opposing and following responses in sensorimotor speech control: Why responses go both ways. *Psychon Bull Rev*, *25*(4), 1458-1467.

Guenther, F. H. (2016). *Neural control of speech*. MIT Press.

Hermes, A., Mertens, J., & Mücke, D. (2018, September 2–6). *Age-related effects on sensorimotor control of speech production* [Paper presentation]. Interspeech 2018, Hybderabad, India.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354), 1213–1216.

Jones, S. A., & Noppeney, U. (2021). Ageing and multisensory integration: A review of the evidence, and a computational perspective. *Cortex*, *138*, 1–23.

Liu, P., Chen, Z., Jones, J. A., Huang, D., & Liu, H. (2011). Auditory feedback control of vocal pitch during sustained vocalization: A cross-sectional study of adult aging. *PLoS One*, *6*(7), e22791.

MacPherson, M. K. (2019). Cognitive load affects speech motor performance differently in older and younger adults. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1258–1277.

Tremblay, P., Sato, M., & Deschamps, I. (2017). Age differences in the motor control of speech: An fMRI study of healthy aging. *Human Brain Mapping*, *38*(5), 2751–2771.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116.

Wieling, M., Rebernik, T., & Jacobi, J. (2023). *SPRAAKLAB: A mobile laboratory for collecting speech production* [Paper presentation]. 20th International Congress of Phonetic Sciences, Prague, Czech Republic.

Onset-cluster production in Mandarin according to sonority profile

Xuejing Chen[1], Pierre Hallé[2], Rachid Ridouane[3]

[1,2,3]Laboratoire de Phonétique et Phonologie (CNRS & U. Sorbonne Nouvelle)

xuejing.chen@sorbonne-nouvelle.fr, pierre.halle@sorbonne-nouvelle.fr, rachid.ridouane@sorbonne-nouvelle.fr

Introduction. The Sonority Sequencing Principle (SSP) might explain a putatively universal preference (part of UG?) for well-formed over ill-formed syllables with respect to onset sonority profile (Clements 1990): the onset of a syllable is well-formed if its sonority profile rises monotonically from the beginning to the nucleus. Speech perception seems to be affected by such sonority-related restrictions. In particular, a series of perceptual studies by Berent and colleagues (e.g., Berent *et al*. 2007; 2008; 2012) showed that *blif* was preferred over *bnif*, *bnif* over *bdif*, and *bdif* over *lbif*, by native listeners of various languages (e.g., English, Korean, Spanish, Mandarin), regardless of whether the onset-clusters involved are permissible or not in the listeners' L1. While numerous studies have focused on the SSP effects in perception, only a few studies have addressed these effects in production and with inconsistent results (Broselow & Finer 1991; Davidson 2000; Redford 2008). In the present study, we explore the possible SSP effects in the production of onset clusters by Mandarin speakers, whose language prohibits clusters in any position. We specifically examine the frequency of schwa occurrence within these clusters, along with their duration. If the SSP influences the production of Mandarin speakers, one would expect a higher incidence of schwa insertions in consonant clusters with a more marked sonority profile, such as fall > plateau > rise.

Methods. Twenty native Mandarin speakers took part in a speech imitation experiment. The participants were presented with a "model" speech stimulus and instructed to faithfully reproduce it, with no time constraints on their responses. The materials to be imitated consisted of 6 pairs of nonwords, either C_1C_2a or $C_1əC_2a$, as detailed in Table 1. The C_1C_2 clusters exhibited rising, plateauing, or falling sonority profiles (referred to as k- or t-pivot for items with /k/ or /t/ in the C_1 position for non-falling profiles or in the C_2 position for falling profiles, respectively). $C_1əC_2a$ items and C_1C_2a items differed solely in the presence of a schwa vowel between C_1 and C_2 in the former. All items were recorded eight times by the third author, a phonetician and native speaker of Tashlhiyt, a language allowing various types of word-initial consonant clusters. The recorded items were scrutinized for the presence or absence of schwas within the C_1C_2 clusters. Two tokens of each item were chosen as models for the experiment. The models for C_1C_2a contained no vocalic material in the inter-consonantal position, while the models for $C_1əC_2a$ included a schwa with a duration ranging from 42 to 102 ms. Each model was presented twice during the experiment, resulting in a total of 48 trials (12 items × 2 models × 2 trials). The trial order was randomized differently for each participant.

Table 1: *C_1C_2a and $C_1əC_2a$ nonword items used in the experiment.*

	Rise		Plateau		Fall	
	k-pivot	t-pivot	k-pivot	t-pivot	k-pivot	t-pivot
C_1C_2a	kla	tla	kpa	tka	lka	lta
$C_1əC_2a$	kəla	təla	kəpa	təka	ləka	ləta

The experiment used SpeechRecorder (Draxler & Jänsch 2004) with an external sound card (Komplete Audio 6 MK2) for subject recording. A total of 960 tokens went into the analysis. We labeled and annotated the data using Praat (Boersma & Weenink 2023). For deciding on the presence/absence of schwa between C_1 and C_2 we followed Ridouane and Fougeron (2011). Three criteria had to be met for a schwa to be labeled: the presence of periodic pulses, an increase in the signal energy at C_1 release, and an interval after C_1 release with formant structure or some energy in the F2/ F3 region characteristic of vowels. We attempted to enforce the classification (presence/absence of schwa), despite the possibility that such strict criteria might have resulted in overlooking some schwas in ambiguous cases. The relative duration of schwa was computed as the ratio between schwa duration and following vowel /a/ duration. We also measured F1 and F2 at the midpoint of schwa in C_1C_2 and $C_1əC_2$; but due to space constraints, the results are omitted from this abstract and will be briefly referenced in the discussion.

Results. Figure 1AB displays the results. In terms of frequency, $C_1əC_2a$ had significantly more schwas than C_1C_2a ($\chi^2(1) = 90.4$, $p < .0001$). Within C_1C_2a, sonority falls yielded more schwas than plateaus ($\chi^2(1) = 25.5$, $p < .0001$), and plateaus

elicited more schwas than rises ($\chi^2(1) = 60.3$, $p < .0001$). In $C_1əC_2a$, sonority falls from C_1 to C_2 also yielded more schwas than rises ($\chi^2(1) = 31.3$, $p < .001$) and plateaus ($\chi^2(1) = 21.8$, $p < .001$). Additionally, relative schwa durations varied significantly based on the sonority profile for both C_1C_2a ($F(2, 295) = 40.1$, $p < 0.001$) and $C_1əC_2a$ ($F(2, 422) = 13.3$, $p < 0.001$): falling sonority in C_1C_2a resulted in longer schwas than rising and plateauing.

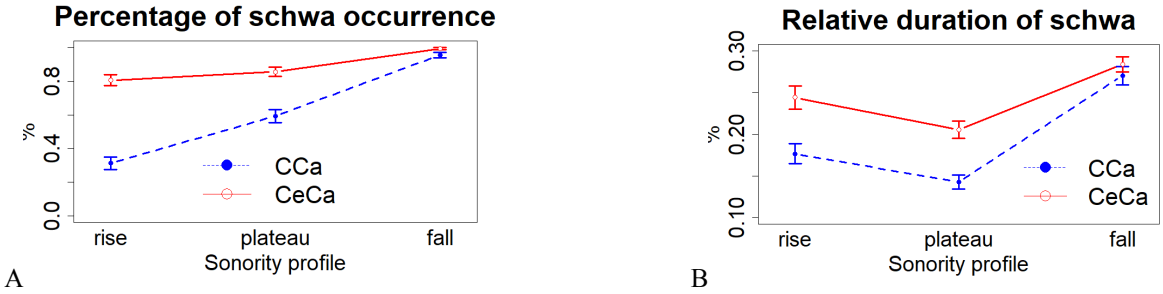


Figure 1AB: *Percentage of occurrence (A) and relative duration (B) of schwas according to sonority profile for C_1C_2a and $C_1əC_2a$ items.*

Discussion. We found clear SSP effects in the production of onset clusters: the more marked the onset cluster, the more likely a schwa was produced in the imitation. Existing literature (e.g., Berent et al., 2007; 2008; 2012) indicates that perceptual repair involving epenthetic vowels typically occurs with nonnative clusters, and this tendency is more pronounced with highly marked clusters. Given that Chinese listeners perceive a schwa within clusters (Zhao & Berent 2016), the data in Figure 1A likely reflects, in part, their perception of the model stimuli. This raises the question of whether our imitation data solely represent perceptual repair of the model stimuli or are modulated by the difficulty of producing consecutive consonants, indexed by the emergence of transitional schwas between consonants. A purely perceptual account would predict similar schwas produced for the C_1C_2 and $C_1əC_2$ models, for example in terms of duration. Yet, schwa duration is shorter for C_1C_2 than $C_1əC_2$ (Fig. 1B). This durational difference suggests that the schwas produced in C_1C_2 are more often unintended, transitional schwas compared to $C_1əC_2$. Although unlikely, an alternative account for the duration data could be that Chinese subjects are sensitive to the durational difference in perception between illusory and real schwas and were able to mimic that difference. Further investigation provides important insight into the nature of schwa produced in C_1C_2 clusters. Analysis of schwa count and durational data in Figure 1AB, along with information on F1 and F2 of schwas in C_1C_2 and $C_1əC_2$, demonstrates that schwa in C_1C_2 with falling sonority onsets closely resembles schwa in $C_1əC_2$ in terms of distribution, duration, and formant structure. For these highly marked onset clusters, the SSP effect is maximal, as Chinese subjects produce similar imitations with "full" schwas for both. The sonority profiles for the other two categories also align with SSP, showing the lowest proportion (~0.3) of produced (and presumably perceived) schwas for /kl, tl/ and an intermediate proportion (~0.5) for /kp, tk/. Although schwa durations are similar for rising and plateau sonority profiles, they are shorter for C_1C_2 than $C_1əC_2$ models and differ in terms of F1 and F2, indicating that some of the schwas produced for C_1C_2 models are targetless and transitional. In sum, we show that the SSP effect is also active in production, and manifests itself in the frequency and the nature of the epenthetic schwa produced in C_1C_2 onsets: part of the time transitional in rising and plateau profile C_1C_2s and almost always full, intended schwas for falling profile C_1C_2s.

References

Berent, I., Lennertz, T., Jun, J., Moreno, M. A., & Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences, 105*(14), 5321–5325.

Berent, I., Lennertz, T., & Rosselli, M. (2012). Universal phonological restrictions and language specific repairs: Evidence from Spanish. *The Mental Lexicon, 13*, 275–305.

Berent, I., Steriade, D., Lennertz, T., & Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition, 104*(3), 591–630.

Boersma, P & Weenink, D. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.4.01, http://www.praat.org.

Broselow, E., & Finer, D. (1991). Parameter setting in second language phonology and syntax. *Second Language Research*, 7, 35–59.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In M. Beckman (Ed.), *Papers in laboratory phonology I: Between the grammar and physics of speech*, 282–333. Cambridge: Cambridge University Press.

Davidson, L. (2000). Experimentally uncovering hidden strata in English phonology. In L. Gleitman & A.Joshi (Eds.), *Proceedings of the 22nd annual conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Draxler, C., & Jänsch, K. (2004). SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software. *International Conference on Language Resources and Evaluation*.

Redford, M. A. (2008). Production constraints on learning novel onset phonotactics. *Cognition, 107*(3), 785– 816.

Ridouane, R. & Fougeron, C. (2011). Schwa elements in Tashlhiyt word-initial clusters. *Laboratory Phonology, 2*(2), 275-30.

Zhao, X. & Berent, I. (2016). Universal restrictions on syllable structure: Evidence from mandarin chinese. *Journal of psycholinguistic research, 45*(4), 795–811.

`praatpicture`: A library for making flexible Praat Picture-style figures in R

Rasmus Puggaard-Rode

Institute for Phonetics and Speech Processing, LMU Munich
r.puggaard@phonetik.uni-muenchen.de

Introduction. The plotting utility available through Praat (Boersma and Weenink 2023), usually accessed through the Praat Picture window of the graphical user interface (GUI), is ubiquitous in phonetics. Praat Picture is a flexible tool which can produce a wide variety of figures, although its most common application is probably plotting one or more acoustic signals which are time-aligned with annotations written in the `.TextGrid` file format; indeed, Praat Picture is undoubtedly the most widely used method for producing this very common style of figure. Praat Picture can either be used with the GUI, which limits the tool's flexibility quite a bit, or with scripts written in Praat's specialized custom scripting language.
The software environment R (R Core Team 2023), which is much more general-purpose than Praat, is used by many phoneticians for a big portion of their processing and analysis pipeline, and increasingly also for preparing manuscripts and presentations using the RMarkdown and Quarto formats (Xie 2015). This presents a need for a similarly flexible plotting utility that can visualize acoustic signals with time-aligned annotations in R, allowing phoneticians to keep as much as possible of their workflow in one software environment.
This paper introduces an R library, `praatpicture`, which aims to fill this gap. The purpose of `praatpicture` is to produce figures of acoustic signals with time-aligned annotations that by default resemble their counterparts in Praat as much as possible, and to allow for at least the same degree of flexibility as plotting in Praat, while relying on signal processing tools that are already available in R. `praatpicture` relies on base R graphics tools, which presents some advantages over Praat, including the ability to resize figures dynamically (i.e. without regenerating figures with new size parameters), and the ability to use any font available to the system. Version 0.6.0 of `praatpicture` is currently available from GitHub (`https://github.com/rpuggaardrode/praatpicture`).

Usage and options. The core function of the library is `praatpicture()`, which only takes one obligatory argument, `sound`, giving the name of a sound file with the `.wav` extension. Calling `praatpicture()` with just one argument will produce a very common figure format: a waveform, a spectrogram, and annotations with dotted lines in the various figure components indicating the locations of annotation boundaries (see the left panel of **Figure 1**). (This assumes that there is a file with the `.TextGrid` extension and the same base name as the `.wav` file in the same directory, but the `make_TextGrid()` function also allows users to create time-aligned annotations interactively in R.) In the following, I give a brief and incomplete overview of the options available to users of the package, some of which are visualized with accompanying code in the right panel of **Figure 1**; argument names which cooccur in the text and in **Figure 1** are given in parentheses in the text and are bolded in the figure.
The user can control the size of individual plot components (`proportion`) and which portion of a sound file to plot (`start`, `end`). Using Praat's terminology, the user can control which annotation tiers to plot (`tg_tiers`) which boundaries to show throughout all figure components (`tg_focusTier`), and the appearance of these boundaries (`tg_focusTierLineType`). In addition to waveforms, spectrograms, and annotations, `praatpicture` can also plot pitch tracks, formant tracks, and intensity tracks (`frames`). The user can control how derived signals are generated – i.e., which window shape and size to use, dynamic range, pitch floor and ceiling, how many formants to calculate, etc. – and how they are visualized – i.e., which frequency range to show (e.g. `pitch_freqRange`), whether pitch and formants should be 'speckled' or 'drawn' (e.g. `pitch_plotType`), which frequency scale to use (e.g. `pitch_scale`), which colors should be used for plotting individual plot components, etc. – with largely the same arguments as those available in Praat. Several options are available for highlighting parts of a figure (e.g. `draw_rectangle`, `annotate`).
A sister function to `praatpicture()`, called `emupicture()`, is available for users of the EMU Speech Database Management System (Winkelmann, Harrington, and Jänsch 2017) who wish to plot annotated signal data directly from an EMU database. No such plotting utility has previously been available. The library also offers the function

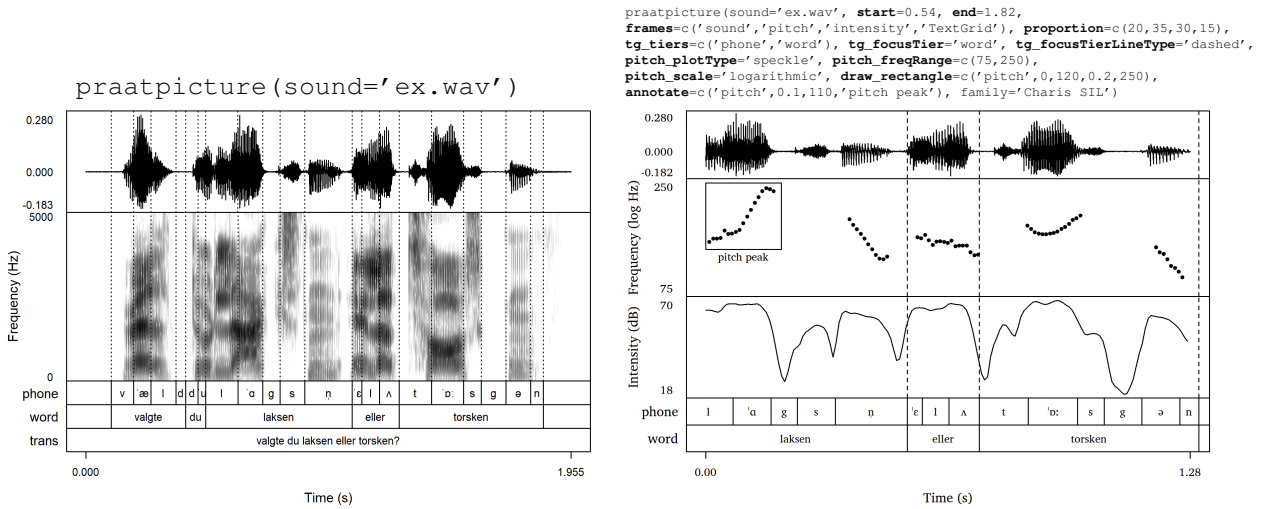


Figure 1: *Two examples of figures generated with* `praatpicture` *and the code used to generate them.*

`talking_praatpicture()` for creating video files of figures with embedded audio, and `praatanimation()` for easily creating `praatpicture()`-based animations.

Implementation. Spectrograms are generated in R using the `phonTools` package (Barreda 2023). Other derived signals are generated using the `wrassp` package in R (Winkelmann, Bombien, et al. 2023). Pitch is calculated using the `ksvF0()` function, formants are calculated using the `forest()` function, and root-mean-squared intensity is calculated using the `rmsana()` function. In all cases, default parameters are set to emulate those in Praat as much as possible, such as e.g. using Gaussian-like window shapes across the board. When results still differ, it is because the underlying algorithms are not identical.
`.TextGrid` files are read into R using the `rPraat` package (Bořil and Skarnitzl 2016), optionally converting to Praat's special character formatting using a custom script. It is also possible to plot pitch, formant, and intensity with `praatpicture()` using values calculated in Praat, if the signals are saved from Praat using the same base file name as the `.wav` file in the same directory; these are then also read into R using `rPraat`. Alternatively, any other software can be used to calculate these signals, as long as they are stored in the Simple Signal File Format (SSFF).

Conclusion. `praatpicture` provides an opportunity for phoneticians who use R (and potentially EMU-SDMS) to keep more of their workflow in R, by allowing users to make familiar-looking figures in a general-purpose software environment without necessarily relying on the plotting and signal processing tools in Praat. Using base R graphics tools to produce these figures arguably has a number of advantages in terms of flexilibity. `praatpicture` currently has most of the same options as Praat does in terms of producing figures with time-aligned acoustic signals and annotations. The library is still in development, so existing features will be augmented and more features will be added over time.

References.

Barreda, Santiago (2023). "`phonTools`. Tools for phonetic and acoustic analyses". (Version 0.2–2.2). URL: `https://CRAN.R-project.org/package=phonTools`.

Boersma, Paul and David Weenink (2023). "Praat. Doing phonetics by computer". (Version 6.4.01). URL: `https://fon.hum.uva.nl/praat/`.

Bořil, Tomáš and Radek Skarnitzl (2016). "Tools `rPraat` and `mPraat`. Interfacing phonetic analyses with signal processing". In: *Text, speech, and dialogue*. Ed. by Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala. Cham: Springer, pp. 367–374. DOI: `10.1007/978-3-319-45510-5_42`.

R Core Team (2023). "R. A language and environment for statistical computing". (Version 4.3.2). URL: `https://R-project.org`.

Winkelmann, Raphael, Lasse Bombien, Michel Scheffers, and Markus Jochim (2023). "`wrassp`. Interface to the ASSP library". (Version 1.0.4). URL: `https://CRAN.R-project.org/package=wrassp`.

Winkelmann, Raphael, Jonathan Harrington, and Klaus Jänsch (2017). "EMU-SDMS. Advanced speech database management and analysis in R". In: *Computer Speech & Language* 45, pp. 392–410. DOI: `10.1016/j.csl.2017.01.002`.

Xie, Yihui (2015). *Dynamic documents with R and* `knitr`. Boca Raton: CRC Press.

Tongue root movement in Hungarian intervocalic alveolar obstruents

Gráczi, Tekla Etelka[1,2,4], Juhász, Kornélia[1,4], Csényi, Péter[1,2], Csapó, Tamás Gábor†[3,4], Deme, Andrea[2,4], Markó, Alexandra[1,4]

[1]*Hungarian Research Centre for Linguistics*
[2]*Eötvös Loránd University*
[3]*Budapest University of Technology and Economics*
[4]*MTA–ELTE Momentum Lingual Articulation Research Group*
graczi.tekla.etelka@nytud.hun-ren.hu, juhasz.kornelia@nytud.hun-ren.hu,
csenyi.peter@nytud.hun-ren.hu, deme.andrea@btk.elte.hu,
marko.alexandra.phd@gmail.com

Introduction. Voiced obstruents are associated with advanced tongue root (TR) position compared to their voiceless counterparts for initiating or maintaining phonation (e.g. Westbury 1983; Proctor, Shadle, and Iskarous 2010). Although the TR-position difference was found in plosives, affricates, and fricatives, as well (e.g. Narayanan, Alwan, and Haker 1995; Ahn 2015; Coretta 2020), considerable within-speaker variation was found in its presence in American English fricatives (Ahn and Davidson 2016) and Hungarian intervocalic /z/ and /s/ (Gráczi et al. 2021). Ahn and Davidson (2016) explain their results by a possible wider obstacle allowing for less need of TR-displacement, but the Hungarian data did not strengthen this assumption (Gráczi et al. 2021). Comparing the results across the cited studies, the question of the role of the vowel context also arises: The larger pharyngeal volume in high vowels compared to low ones (Baer et al. 1991) might result in a higher resistance of high vowels against further pharynx enlargement, allowing for later or no TR-advancement in the following voiced obstruents than the lower vowels. The present study raised the following question based on the patterns described above: How do the vocalic context (front vs. back V) and the manner of articulation (plosive/fricative/affricate: MoA) interact in terms of the tongue root advancement in Hungarian voiced unaspirated and voiceless unaspirated obstruent pairs?

Methods. Hungarian alveolar obstruents (/d/, /t/, /z/, /s/, /ʣ/, /ʦ/) were embedded in /l/V_V/l/ nonsense words, where the vowels were identical either /i/ or /ɒ/. We have to note that intervocalic /ʣ/ in Hungarian surfaces as a long consonant in real words, while varies in non-sense words. The 12 nonsense words were read aloud by 5 native speakers of Hungarian 5 times each in random order among further nonsense words. The recordings were carried out by AAA ultrasound device, software (83 images/s), and an omnidirectional condenser microphone. The ultrasound probe was fixed by a helmet. The segment boundaries were automatically labeled (Mihajlik et al. 2009) and manually corrected in Praat (Boersma and Weenink 2023) based on the F_2 offset and onset. The tongue contours were drawn semi-automatically and manually corrected in AAA. The tongue root movement was traced by selecting the spline showing the largest position variance at the TR-region (see Coretta 2020). The consonant duration was normalised to an interval between 1 and 2. The TR-position was Z-normalized by speaker and multiplied by -1 so the higher number reflects more advanced TR. The preliminary data of three female speakers are shown, the manual correction of the remaining 2 speakers' tongue contours is being carried out. Generalized additive mixed models (GAMMs) were run in R (R Core Team 2022) with mgcv v1.8-26 following: Sóskuthy (2021) (Wood 2017; van Rij et al. 2022). Random slopes by the speaker and random smooths were included besides the ordered factors of MoA, voicing and their interaction both as parametric and smooth terms.

Results. The results for the **parametric terms** of GAMMs for the /i/ context only showed a significant difference in the general (mean) position of TR across the three MoAs, while the results for /ɒ/-context indicate that the averaged TR-position is significantly different for the factor voicing, and MoA as main effects. No difference was found in the **smooth terms** in any of the vowel contexts, which means that the shape of the TR-movement trajectories along the consonant is similar regardless of MoA and/or voicing. The **smooth-** (**Figure 1**) **and difference-estimates** indicate that the TR-

position does not show any significant difference along the duration of the consonants in /i/-contexts, while in the case of /ɒ/-context, the TR is more advanced around the middle of the duration for voiced plosives and affricates, and along the entire duration of the voiced fricatives. The duration and duration ratio results of /ʤ/ were different than the other /z/ and /d/ but the voiced part ratio and tongue root movement varied with these.

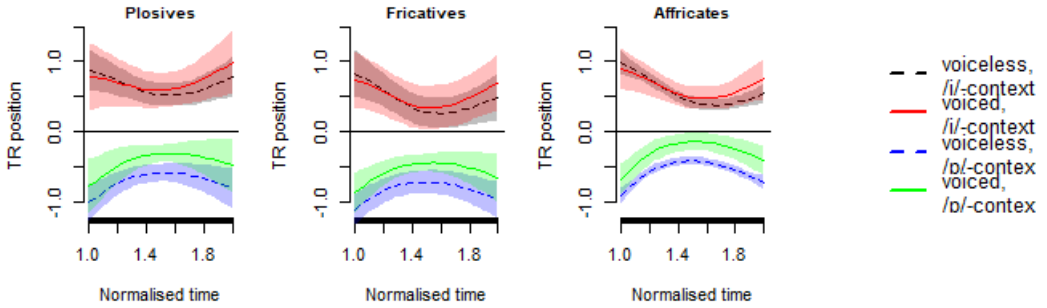


Figure 1: *The estimated TR-movement within the consonant duration (GAMMs were run separately for the vowels, 3 speakers' data).*

Discussion. The preliminary results of the present study indicate that the TR-advancement associated with the voiced obstruents compared to their voiceless counterparts is context-dependent, and the timing of the advancement when present varies across the manner of articulations.

References.

Ahn, S. (2015). "Tongue root contributions to voicing in utterance-initial stops in American English". In: *Proceedings of Meetings on Acoustics*. Vol. 25. 1. AIP Publishing.

Ahn, S. and L. Davidson (2016). "Tongue root positioning in English voiced obstruents: Effects of manner and vowel context". In: *The Journal of the Acoustic Society of America* 140, p. 3221.

Baer, T., J. C. Gore, L. C. Gracco, and P. W. Nye (1991). "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels". In: *The Journal of the Acoustic Society of America* 90.2, pp. 799–828.

Boersma, P. and D. Weenink (2023). *Praat: doing phonetics by computer*. URL: `http://www.praat.org/`.

Coretta, S. (2020). "Longer vowel duration correlates with greater tongue root advancement at vowel offset: Acoustic and articulatory data from Italian and Polish". In: *The Journal of the Acoustical Society of America* 147.1, pp. 245–259.

Grászi, T. E., T. G. Csapó, A. Deme, K. Juhász, and A- Markó (2021). "Tongue root position in VC sequences with regard to the phonetic realization of obstruent voicing: A preliminary study on Hungarian". In: *Proc. of the 12th International Seminar on Speech Production*, pp. 198–201.

Mihajlik, P., Z. Tüske, B. Tarján, B. Németh, and T. Fegyó (2009). "Improved recognition of spontaneous Hungarian speech—Morphological and acoustic modeling techniques for a less resourced task". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.6, pp. 1588–1600.

Narayanan, S. S., A. A. Alwan, and K. Haker (1995). "An articulatory study of fricative consonants using magnetic resonance imaging". In: *The journaltitle of the Acoustical Society of America* 98.3, pp. 1325–1347.

Proctor, M. I., C. H. Shadle, and K. Iskarous (2010). "Pharyngeal articulation in the production of voiced and voiceless fricatives". In: *The journaltitle of the Acoustical Society of America* 127.3, pp. 1507–1518.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. URL: `https://www.R-project.org/`.

Sóskuthy, M. (2021). "Evaluating generalised additive mixed modelling strategies for dynamic speech analysis". In: *Journal of Phonetics* 84, p. 101017.

van Rij, J., M- Wieling, R. H- Baayen, and H. van Rijn (2022). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. R package version 2.4.1.

Westbury, J. R. (1983). "Enlargement of the supraglottal cavity and its relation to stop consonant voicing". In: *The journaltitle of the Acoustical Society of America* 73.4, pp. 1322–1336.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. second. Chapman and Hall/CRC.

Acoustic correlates of the nasal vs. plosive quantity contrast in Hungarian

Tilda Neuberger

HUN-REN Hungarian Research Centre for Linguistics

neuberger.tilda@nytud.hun-ren.hu

Introduction. Length serves as a distinguishing feature between two sets of consonants, namely singletons and geminates, in a variety of languages. Previous research has demonstrated that a range of durational and non-durational acoustic parameters play a role in contributing to the quantity contrast, although the extent of their influence varies across languages (e.g., Al-Tamimi & Khattab 2018; Amano *et al.* 2021; Hermes *et al.* 2020).
Furthermore, the realization of consonant length may vary across consonant types. Different features are expected to contribute to the expression of quantity in obstruent vs. sonorant consonants, given their distinct spectral structures, for instance, their spectral continuity. Listener perception seems to differ depending on the consonant type, with short/long pair discrimination being more challenging in nasals than in obstruents (Kawahara & Pangilinan 2017).
In Hungarian, geminates can occur in all consonant types, including, but not limited to, nasals and plosives. This provides an ideal context for investigating the quantity contrast according to the consonant type.
The aim of this study is to explore the acoustic parameters contributing to the length opposition in Hungarian nasals and plosives. We hypothesize that speakers mark the contrast differently depending on the consonant type. Given the challenge spectral continuity poses to perceiving length contrast, it is plausible that speakers use the durational parameter more robustly in expressing nasal quantity contrast than plosive quantity contrast or enhance the nasal quantity contrast with additional secondary acoustic features.

Methods. Intervocalic nasal /n ɲ/ and plosive /t k/ singletons and geminates (N = 400) were collected from the spontaneous speech of 20 monolingual Hungarian-speaking adults using the BEA database (Neuberger et al., 2014). Various acoustic parameters were measured by means of Praat (Boersma & Weenink 2020), including absolute and relative durations of the target consonants and their surrounding vowels. Linear mixed-effects models were constructed using R (Bates *et al.* 2014) for each acoustic parameter to investigate the effect of quantity (singleton vs. geminate), consonant type (nasal vs. plosive) and their interaction. Additionally, decision trees were employed to identify the most important features in distinguishing the two phonological length categories in nasals and plosives.

Results. Preliminary results indicated significant differences in the consonant duration between singletons and geminates in both nasals and plosives (see Figure 1). A significant interaction between consonant quantity (S vs. G) and consonant type (nasal vs. plosive) on consonant duration was observed. The G/S ratio was significantly higher for nasals compared to plosives, indicating a more distinct contrast in nasals. Closure duration proved to be the most important acoustic correlate of consonant length in plosives. Acoustic results also showed that the duration of the surrounding vowels helps distinguish the two phonological categories, with a greater contribution shown for nasals.

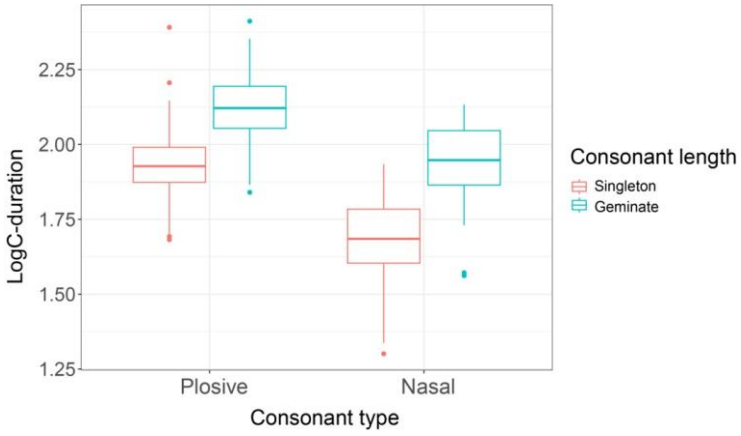


Figure 1: *Consonant duration (log-transformed) as a function of consonant length and consonant type*

Discussion. Our findings suggest that the expression of the quantity contrast in nasals requires more robust time adjustments than in plosives. This reflects the previous finding (see Kawahara & Pangilinan 2017) that listeners have more difficulty distinguishing the length contrast in spectrally continuous sounds (like nasals), and therefore speakers put more effort into their production to ensure successful comprehension.
The results of this study contribute to a more accurate description of the phonetic realization of phonological length in Hungarian, and may bring us closer to understanding the preferential hierarchy of geminate occurrences across languages, namely that obstruent geminates are more likely to occur in a language than nasal geminates.

References

Al-Tamimi, J., & Khattab, G. (2018). Acoustic correlates of the voicing contrast in Lebanese Arabic singleton and geminate stops. *Journal of Phonetics,* 71, 306-325.

Amano, S., Kondo, M., & Yamakawa, K. (2021). Predicting and classifying Japanese singleton and geminate consonants using logarithmic duration. *The Journal of the Acoustical Society of America,* 150(3), 1830-1843.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Boersma P, Weenink D (2020) Praat: Doing phonetics by computer (Version 6.1.30) [Computer Program]. http://www.praat.org

Hermes, A., Tilsen, S., & Ridouane, R. (2020). Cross-linguistic timing contrast in geminates: A rate-independent perspective. In *Proceedings of the* 12th *International Seminar on Speech Production (ISSP2020).* 52-55.

Kawahara, S., & Pangilinan, M. (2017). Spectral continuity, amplitude changes, and perception of length contrasts. In Kubozono, H. (Ed.): The phonetics and phonology of geminate consonants. Oxford: Oxford University Press, 13-33.

Neuberger, T., Gyarmathy, D., Gráczi, T. E., Horváth, V., Gósy, M., & Beke, A. (2014). Development of a large spontaneous speech database of agglutinative Hungarian language. In *Proceedings of Text, Speech and Dialogue (TSD2014).* Springer International Publishing. 424-431.

Speech and language abilities associated with regional corpus callosum development in children who stutter

Fiona Höbler[1], Emily O. Garnett[1], Yanni Liu[1], Ho Ming Chow[2], Soo-Eun Chang[1]

[1]University of Michigan, [2]University of Delaware

fhobler@med.umich.edu, emilyog@med.umich.edu, yanniliu@med.umich.edu, hmc@udel.edu, sooeunc@med.umich.edu

Introduction. As the major commissural tract interconnecting the cerebral hemispheres, the corpus callosum supports both the interhemispheric transfer of information as well as hemispheric lateralization of specialized functions such as speech and language. Increased region-specific volumes of the corpus callosum have been linked to better expressive language abilities and verbal fluency, as well as increased left and decreased right hemispheric lateralization of language connectivity in children (Bartha-Doering et al., 2021). Among adults, however, increased volume of the corpus callosum has been associated with reduced lateralization of speech processes (Lassonde et al., 1990). Both age- and sex-based distinctions have been established in regional callosal development (Luders et al., 2010), as well as morphological differences across several neurodevelopmental conditions (Paul, 2011). In developmental stuttering, a larger overall area of the corpus callosum as well as increased anterior white matter volume (WMV) were reported for adults who stutter relative to adults who do not stutter (Choo et al., 2011). Regional differences in callosal size or WMV have not been found in children who stutter (CWS) when measured at school age (Choo et al., 2012). However, a reduced overall growth rate of WMV in the corpus callosum was found in preschool CWS (Chow et al., 2023). When compared to children who do not stutter (CNS), decreased regional white matter integrity (as measured in fractional anisotropy) has also been reported in CWS (Chang et al., 2015). In addition, atypical laterality in language related function and structure among persons who stutter was reported. Adults with persistent developmental stuttering showed increased rightward activation during speech production (De Nil et al., 2000; Fox et al. 1996), larger surface area of right Perisylvian language regions (Cykowski et al., 2000), as well as reduced prefrontal and occipital volumes associated with lower language abilities (Foundas et al., 2003). Yet, how neurodevelopmental differences in the structure of this interhemispheric commissure relate to language abilities in developmental stuttering has not been examined in children. In this study, we investigated whether expressive language abilities and speech sound articulation are associated with differences in WMV of the corpus callosum in preschool- and school-aged CWS and CNS. Based on previous research (Bartha-Doering et al., 2021), we hypothesized that better language abilities would be associated with larger anterior and posterior volume, with potentially differential mediation of sex and age-group effects between CWS and CNS in these regions.

Methods. Participants in this study were part of a larger longitudinal brain imaging investigation of developmental stuttering, in which children were scanned up to 4 times (1 visit per year). The current study included 74 CWS (28 female) and 75 CNS (36 female) for whom 405 scans with useable structural data, as well as expressive language and speech sound accuracy scores were available. Speech and language assessment included the Expressive Vocabulary Test Third Edition (EVT-3) or Expressive Language subtest as part of the Clinical Evaluation of Language Fundamentals Preschool-2 (CELF P-2), along with speech sound accuracy on the Goldman-Fristoe Test of Articulation (GFTA-2). All participants also completed the Wechsler Abbreviated Scale of Intelligence (WASI-II) or Wechsler Preschool & Primary Scale of Intelligence (WPPSI-IV). Age-based standard scores for each speech and language test were used as individual expressive language and speech sound accuracy measures in the analyses described below. All MRI scans were acquired on a GE 3T Signa HDx MR scanner with an 8-channel head coil. During each session, 180 T1-weighted 1-mm3 isotropic volumetric inversion recovery fast spoiled gradient-recalled images (3D IRFSPGR) (10 min scan time), with CSF suppressed, were obtained to cover the whole brain. Volumetric measures of the corpus callosum were derived using FreeSurfer 5.3.0, which automatically segmented individual anatomical images into five regions, equally spaced in distance along the long axis: anterior, mid-anterior, central, mid-posterior and posterior. Analyses were conducted within groups of preschool- (3 – 5;11 years of age) and school-aged (6 – 12;11 years of age) CWS and CNS. Differences on speech and language measures were first investigated using linear effects modeling, with fixed effects of group and sex included. To investigate volumetric differences between the groups, separate linear mixed effects models were used for each of the five sub-regions, with the predictive influence of expressive language and speech sound accuracy analyzed separately. To investigate callosal development, main effects of age and sex were modeled to allow for interactions between group, age, and sex, with total brain volume added as effect of no interest. All models also controlled for IQ, socioeconomic status (SES) and individual variability.

Results. Among school-aged children, CWS produced lower scores on measures of speech sound accuracy and IQ than CNS, while preschool CWS produced lower scores on measures of expressive language than their age-matched peers. Across the five subregions of the corpus callosum, CWS and CNS were found to differ in growth of WMV in the mid-

anterior segment only. Specifically, among school-aged male participants, CWS had reduced growth rate of mid-anterior volume when compared to age- and sex-matched CNS (CWS x Age in months: $\beta = 0.84$, $SE = 0.40$, $t = 2.10$, $p = .04$), which was non-significant following correction. When investigating the relationship between language abilities and WMV of the corpus callosum among preschool- and school-aged CWS and CNS, results indicated that expressive language abilities were associated with WMV in the mid-posterior as well as mid-anterior regions differentially. At preschool age, higher expressive language scores among male CWS were associated with greater WMV in the mid-posterior (Expressive language x CWS: $\beta = 4.44$, SE = 1.38, t = 3.22, p = .002) and mid-anterior corpus callosum (Expressive language x CWS: $\beta = 3.53$, SE = 1.32, t = 2.67, p = .01), while higher expressive language scores among male CNS, at preschool as well as school age, reflected lower WMV in these regions. Among female participants, higher scores in speech sound accuracy among CWS were associated with reduced WMV in the mid-posterior corpus callosum, while better speech sound accuracy was associated with increases in WMV for CNS, at both preschool (GFTA x CWS: $\beta = -4.14$, SE = 1.53, t = -2.71, p = .01) and school age (GFTA x CWS: $\beta = -4.84$, SE = 1.66, t = -2.91, p = .005). No significant effects of speech sound accuracy were found among male participants.

Discussion. The results of the current study bring to light significant age and sex-specific differences between CWS and CNS in regional corpus callosum development, as well as the differential influence of speech and language abilities between boys and girls. At preschool age, better expressive language abilities among boys who stutter were associated with larger WMV in the mid-anterior and mid-posterior corpus callosum; however, school-aged boys who stutter tended to show decreases with age in WMV in the mid-anterior region. While the mid-posterior region has been associated with the interhemispheric transfer of sensory information, the mid-anterior supports excitatory and inhibitory interhemispheric communication between the premotor and supplementary motor regions (Hofer & Frahm, 2006). Previous research found decreased cortical thickness in premotor and primary motor areas of children experiencing persistent developmental stuttering, compared to children who recover and children who do not stutter (Garnett et al., 2018). Differences in mid-anterior volume reflect similar findings among adults with persistent developmental stuttering (Choo et al., 2011), and may indicate important neurodevelopmental deficiencies in interhemispheric connectivity that supports motor control which also relate to stuttering persistence. Better speech sound accuracy among girls who stutter was associated with decreased volume in the mid-posterior corpus callosum, a region reported to correlate positively with verbal fluency (Nosarti et al. 2004) as well as with connectivity between left language-related regions including the left inferior frontal cortex, insular cortex, and precentral gyrus (Bartha-Doering et al., 2021). Taken together, these findings point towards the significant role of regional callosal development in supporting processes of motor functioning as well as speech and language development during childhood.

References

Bartha-Doering, L., Kollndorfer, K., Schwartz, E., Fischmeister, F. P. S., Alexopoulos, J., Langs, G., ... & Seidl, R. (2021). The role of the corpus callosum in language network connectivity in children. *Developmental Science*, 24(2), e13031.

Chang, S-E., Zhu, D.C., Choo, A., Angstadt, M. (2015). White matter neuroanatomical differences in children who stutter. *Brain*. 2015 Mar;138(Pt 3):694-711.

Choo, A.L., Chang, S.-E., Zengin, H., Ambrose, N.G., & Loucks,T. (2012). Corpus callosum morphology in children who stutter. *Journal of Communication Disorders*. 45(4), 279-289.

Choo, A.L., Kraft, S.J., Olivero, W., Ambrose, N.G., Sharma, H., Chang, S. & Loucks,T. (2011). Corpus Callosum differences associated with persistent stuttering in adults. *Journal of Communication Disorders*. 44(4), 470-477.

Chow, H. M., Garnett, E. O., Koenraads, S. P., & Chang, S. E. (2023). Brain developmental trajectories associated with childhood stuttering persistence and recovery. *Developmental Cognitive Neuroscience*, 60, 101224.

Cykowski, M. D., Kochunov, P. V., Ingham, R. J., Ingham, J. C., Mangin, J. F., Riviere, D., ... & Fox, P. T. (2008). Perisylvian sulcal morphology and cerebral asymmetry patterns in adults who stutter. *Cerebral Cortex*, 18(3), 571-583.

De Nil, L. F., Kroll, R. M., Kapur, S., & Houle, S. (2000). A positron emission tomography study of silent and oral single word reading in stuttering and nonstuttering adults. *Journal of Speech, Language, and Hearing Research*, 43(4), 1038-1053.

Foundas, A. L., Corey, D. M., Angeles, V., Bollich, A. M., Crabtree–Hartman, E., & Heilman, K. M. (2003). Atypical cerebral laterality in adults with persistent developmental stuttering. *Neurology*, 61(10), 1378-1385.

Fox, P. T., Ingham, R. J., Ingham, J. C., Hirsch, T. B., Downs, J. H., Martin, C., ... & Lancaster, J. L. (1996). A PET study of the neural systems of stuttering. *Nature*, 382(6587), 158-162.

Garnett, E. O., Chow, H. M., Nieto-Castañón, A., Tourville, J. A., Guenther, F. H., & Chang, S. E. (2018). Anomalous morphology in left hemisphere motor and premotor cortex of children who stutter. *Brain*, 141(9), 2670-2684.

Hofer, S., & Frahm, J. (2006). Topography of the human corpus callosum revisited—comprehensive fiber tractography using diffusion tensor magnetic resonance imaging. *Neuroimage*, 32(3), 989-994.

Lassonde, M., Bryden, M. P., & Demers, P. (1990). The corpus callosum and cerebral speech lateralization. *Brain and Language*, 38(2), 195-206.

Luders, E., Thompson, P. M., & Toga, A. W. (2010). The development of the corpus callosum in the healthy human brain. *Journal of Neuroscience*, 30(33), 10985-10990.

Nosarti, C., Rushe, T. M., Woodruff, P. W., Stewart, A. L., Rifkin, L., & Murray, R. M. (2004). Corpus callosum size and very preterm birth: relationship to neuropsychological outcome. *Brain*, 127(9), 2080-2089.

Paul, L. K. (2011). Developmental malformation of the corpus callosum: a review of typical callosal development and examples of developmental disorders with callosal involvement. *Journal of Neurodevelopmental Disorders*, 3(1), 3-27.

Phoneme monitoring and articulatory suppression in French-speaking adults

Claire Boilley[1,2], Patricia Pires[1], Anne Vilain[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
[2]Univ. Grenoble Alpes, CNRS, LPNC, 38000 Grenoble, France
Claire.boilley@univ-grenoble-alpes.fr, anne.vilain@univ-grenoble-alpes.fr

Introduction. Phonological awareness refers to the ability to detect and manipulate sublexical units in spoken words. Phonological awareness at the phoneme level (phoneme awareness), in particular, has strong links to reading acquisition (Melby-Lervåg, 2012). However, the ability to explicitly segment a word into a string of phonemes seems to only develop after the onset of learning to read in an alphabetical system. We suggest that individuals who have not learned to read yet struggle with this type of task because they use a very concrete strategy, monitoring of articulation, which is not particularly suitable for phonemes, notably because of coarticulation. Mastering the alphabetical principle allows one to use more abstract representations. Arguments for an articulatory strategy stem from the observation that pre-reading children perform better at segmenting VC than CV ; the latter requires a glottal stop after the onset consonant which breaks the natural flow of speech (Geudens et al., 2004). Articulatory traits of individual phonemes also seem to influence how easily they are extracted from words (de Graaff et al., 2011). On the other hand, adults instructed to judge the presence or absence of a given phoneme in words (phoneme monitoring) are hardly disturbed by concurrent articulatory suppression (Wheeldon & Levelt, 1995), suggesting a more abstract approach to phonemic segmentation. We replicate Wheeldon & Levelt's experiment with healthy French adults, adding an auditory interference task to control for the effect of auditory feedback while articulating, and a semantic task to control for the effect of dual tasking.

Methods. Approval for this experiment was obtained from a local ethics committee. We tested 42 students from Grenoble University (18 to 31 years old, F=37, M=5) who were native French speakers, with normal or corrected-to-normal vision and hearing, and no history of speech or language disorder.
The experimental task was a phoneme monitoring task, in which participants looked at an image, and were asked to tell whether a target phoneme was included in the name corresponding to that image, by pressing on a button. This phonological task was performed in three different conditions: (1) the *simple* condition, where no additional task was asked, (2) the *articulatory suppression* condition, in which subjects had to complete the task while repeating the non-word /bakusi/ continuously, starting before the onset of the first item, (3) the *auditory interference* condition, in which the participants performed the task while listening to a recording of /bakusi/ pronounced by a native female speaker and played continuously. In order to control for the effect of dual tasking in each of these three conditions, a semantic task was added, and performed in the same three conditions. The task consisted in looking at images and answering questions about these images (e.g. "Is this an animal from the farm?") by pressing on a button.
Target phonemes in the phonological task were 6 consonants of various phonetic classes with regular spelling: /p/, /t/, /d/, /r/, /l/, /m/. The phoneme /s/, whose spelling is highly variable, was used as a target in a training block to attract participants' attention to the fact that the target was a speech sound, not its spelling. The targets appeared in disyllabic words with a length of 5 segments, controlled for frequency and naming agreement. Each target phoneme appeared once in each of the following positions: word-initial (C1), word-medial after a « simple » CV syllable (C2), word-final (C3), or word-medial after a CVC or CCV syllable (Cplx). Fillers for the phonological task and items for the semantic task were chosen from the same database. We aimed to favor words comparable in length and structure to the carrier words, although this was not a strict criterion. Participants were familiarized with the pictures by viewing and naming (or if needed, repeating from the examiner) each of them once.
Tasks were completed in a fixed order (semantic then phonological), with each including the 3 conditions also in a fixed order (simple, articulatory suppression, and auditory interference). In the phonological task, each condition comprised 2 blocks, with one target phoneme per block, whose identity was counterbalanced across 3 groups of participants so that each phoneme was processed in each position and condition overall. Participants were told at the beginning of each block which phoneme to detect, and instructed to respond by pressing either the left ("no") or right ("yes") arrow key on the computer's keyboard. Each condition in each task started with a training block.

Results. Generalized linear modeling (*glmer* function in R, family= binomial) was used to analyze accuracy of responses to target words (carrier words), with task and condition and their interaction as fixed effects, and participant as random effect (**Figure 1**),. We find a main effect of task and of condition, but no interaction between task and condition. Multiple comparisons (*emmeans* function) reveal that scores are lower under articulatory suppression only (F=3.49, p=.012).

Linear mixed models (*lme* function) on log-transformed reaction times (RTs, correct responses only) show an effect of task (F=769.23, p<.001), an effect of condition (F=8.42, p<.001) and a significant interaction (F=4,29, p = 0.014) (**Figure 1**, right). Multiple comparisons reveal that there is no effect of condition in the phonological task, while there is a facilitating effect of auditory interference in the semantic task Auditory interference is found to reduce RTs in the semantic condition only, as compared to the simple condition (t=4.84, p<.001), and to articulatory suppression (t=2.803, p=.014). As for positional effects, consonants are detected more accurately at word onset (C1) than in any other position across all conditions (p<.001). Analysis of reaction times shows no statistically significant effect of the interaction between position and condition. Surprisingly, word-medial consonants (C2) are found equally fast as word-initial consonants (C1), while other positions take longer to process (p<.01). Multiple comparisons by condition suggest that articulatory suppression makes latencies so variable they no longer show any significant difference between different target positions.

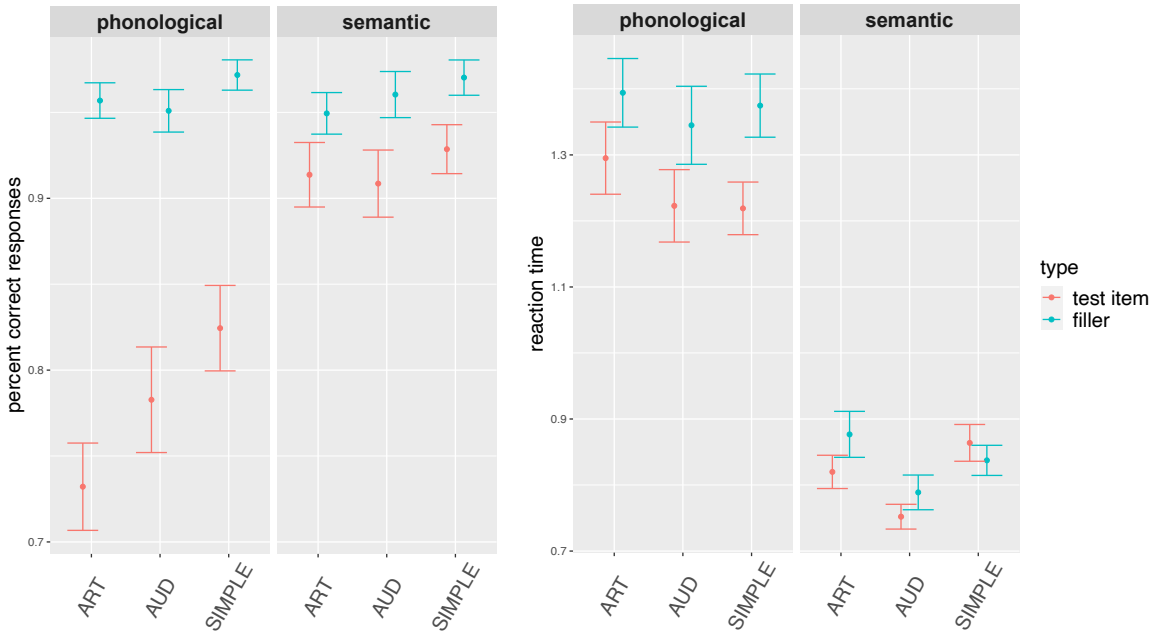


Figure 1: *percent correct responses (left), reaction times in sec. (RT) of correct response* (right), *in the phonological and semantic tasks, in the three conditions (articulatory suppression, auditory interference, simple)*.

Discussion. Articulatory suppression was found to significantly lower accuracy and lengthen reaction times in both the phoneme monitoring task and in the semantic task in our participants. This result suggests that this effect may be explained by the generic additional cognitive cost of articulatory suppression. This is in line with Wheeldon & Levelt's (1995) results, and with our hypothesis that healthy literate adults use abstract phonological representations to segment words. Furthermore, and in contradiction with Wheeldon & Levelt (1995)'s experiment in Dutch, we find no significant difference in how fast subjects access word-medial vs word-initial consonants when the first syllable has a CV structure. This difference might be due to the fact that Wheeldon & Levelt used a list of words with either a simple CV or a more complex first syllable, whereas we separated those two types. Lexical stress patterns may also be at play. In French, unlike in Dutch, lexical stress typically affects the last syllable, possibly making it faster to analyze, although not to the point of counteracting the slowing effect of a complex first syllable.

One limitation to this study is that articulatory suppression may not (always) be a reliable tool to measure reliance on articulatory representations, notably because of subtle temporal strategies. Some participants systematically inserted short breaks, which might have facilitated the processing of items, between each utterance of the nonword, despite instruction not to do so. We may thus need to revise our protocol to better control the temporal alignment of item onset and articulation. Our future goals are first to measure overt articulation times recorded during the familiarization task to analyze the relationship between duration of the first syllable and latency of detection of a consonant at the onset of the second syllable; then, to adapt the task to make it suitable for pre-reading children and young readers.

References

de Graaff, S., Hasselman, F., Verhoeven, L., & Bosman, A. M. T. (2011). Phonemic awareness in Dutch kindergartners : Effects of task, phoneme position, and phoneme class. *Learning and Instruction*, *21*(1), 163-173. https://doi.org/10.1016/j.learninstruc.2010.02.001

Geudens, A., Sandra, D., & Van den Broeck, W. (2004). Segmenting two-phoneme syllables : Developmental differences in relation with early reading skills. *Brain and Language*, *90*(1-3), 338-352. https://doi.org/10.1016/S0093-934X(03)00446-2

Melby-Lervåg, M. (2012). The Relative Predictive Contribution and Causal Role of Phoneme Awareness, Rhyme Awareness and Verbal Short-Term Memory in Reading Skills : A Review. *Scandinavian Journal of Educational Research*, *56*(4), 363-380. https://doi.org/10.1080/00313831.2011.594611

Wheeldon, L. R., & Levelt, W. J. M. (1995). Monitoring the Time Course of Phonological Encoding. *Journal of Memory and Language*, *34*(3), 311-334. https://doi.org/10.1006/jmla.1995.1014

Dimensions of structure and variability in the human vocal tract

Katherine Vaughan-Williams[1], Steven Moran[2,3], Sam Kirkham[1]

[1]*Phonetics Laboratory, Lancaster University*
[2]*Institute of Biology, University of Neuchâtel*
[3]*Department of Anthropology, University of Miami*
kpvaughanwilliams@gmail.com, steven.moran@unine.ch, s.kirkham@lancaster.ac.uk

Introduction.
A defining characteristic of the human vocal tract is a complex dynamic between structure and variability. Across a population we observe considerable variability in vocal tract dimensions, ranging from sexual dimorphism in vocal tract length (Fitch and Giedd 1999) and oral cavity length (Fant 1966), to individual differences in the hard palate (Lammert, Proctor, and Narayanan 2013). At the same time, variation in one dimension is rarely independent of other dimensions and, in many cases, correlates with other aspects of the body, such as speaker height and weight (Stone et al. 2018). Are some of these relationships, however, more variable than others? Do some vocal tract dimensions always scale together uniformly, or do they show a more non-linear relationship in different areas of a parameter range? In this study, we report a data-driven investigation into the relationship between vocal tract dimensions. We first identify the primary dimensions of variation in the vocal tract, followed by an analysis of multi-dimensional interactions between parameters.

Methods.
We use Magnetic Resonance Imaging data of the vocal tract, taken from 69 speakers in the USC Speech MRI Database (Lim et al. 2021). Measurements were extracted by hand from the midsagittal vocal tract using ImageJ (Schneider, Rasband, and Eliceiri 2012), based on a single representative rest posture for each speaker. Measurements include vocal tract length, palate length, palate height, tongue length, and tongue area, as well as each speaker's height, weight and body-mass index (BMI). Our analysis is twofold: (1) what are the primary dimensions of variability? (2) what are the relationships between vocal tract parameters? We addressed (1) by submitting all measures to Principal Components Analysis. We find that two components capture 79.5% of the variance and we explore how each vocal tract measurement corresponds with those dimensions to understand the primary axes of variation in the data. The second analysis then aims to better understand the precise relationship between vocal tract parameters. We fitted a series of conditional inference trees to each variable in the data set, with all other variables used as predictor variables. Conditional inference trees are a class of regression models using binary recursive partitioning. This tests the relationship between each predictor variable and the outcome variable: the most important variable in predicting the outcome is chosen as the top variable in the tree, and then another significance test is carried out on the other predictor variables within each level of the top-most predictor. This process is repeated recursively, unless all variables have been exhausted. We visualise the models as in Figure 1, where the predictor variables are ordered from top-to-bottom in terms of importance, with the boxplots representing terminal nodes that correspond to the distribution of data points within that combination of variables.

Results.
The PCA results show that PC1 captures variation across vocal tract length and tongue length/area, while PC2 captures variation in palate height, which is highly independent of the vocal tract/tongue measures. Palate length is equally weighted across both dimensions, showing its interaction with both palate height and vocal tract/tongue length. K-means clustering on these PC values reveals two separable clusters, which highly correlate with speaker sex. The conditional inference trees expose more precise relationships between parameters. Figure 1 shows such a model with vocal tract length as the outcome variable and all other variables as potential predictors. The model finds five distributions in the data, based on the interaction between three predictor variables. Speaker sex is unsurprisingly the strongest predictor of vocal tract length. Within male speakers, there is one split in the distribution, such that speakers with a smaller tongue area are more likely to have a smaller vocal tract. Within female speakers, a similar split occurs, but for tongue length.

Finally, within female speakers with a shorter tongue, there is a further split based on small differences in tongue area. The other variables show no significant association with vocal tract length. Models fitted to other variables also reveal a high degree of association between tongue length and palate length. We additionally find that speaker sex can be clearly (but not perfectly) predicted from a speaker's vocal tract length, but none of the body measurements (height, weight, BMI) are strong predictors of any of the vocal tract measures.

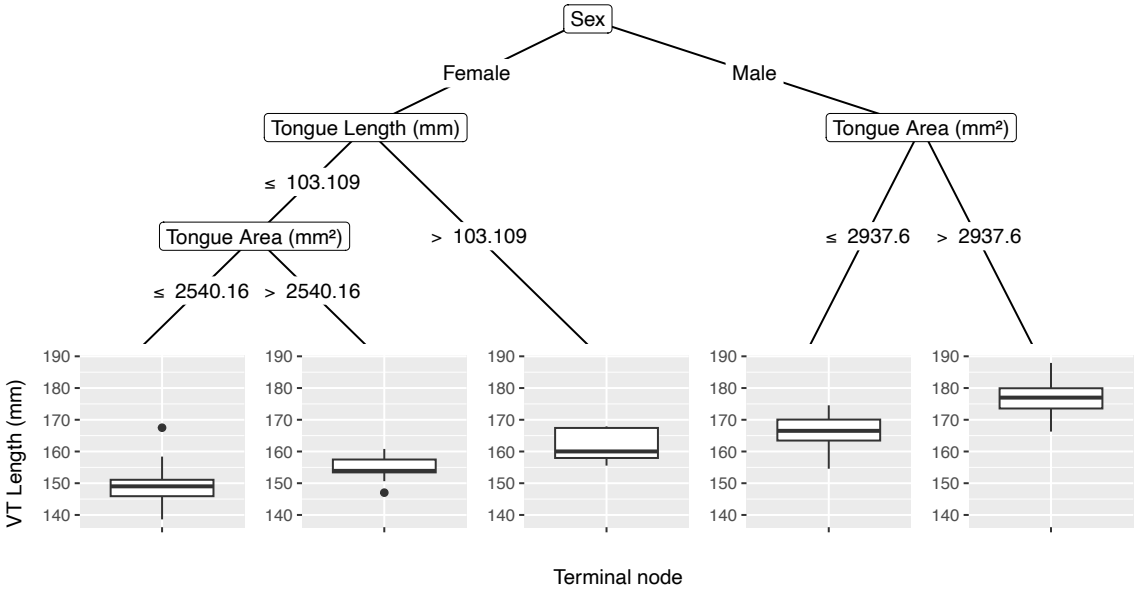


Figure 1: *Conditional inference tree fitted to vocal tract length measurements. Predictors that do not appear on the plot are not significant predictors of vocal tract length in the model.*

Discussion.

Our results suggest that some dimensions of the vocal tract show very clear patterns, such as sexual dimorphism in vocal tract length. Elsewhere, however, relationships between vocal tract parameters are more complex and often non-linear. Specifically, there may be differential effects of a given measure (e.g. tongue length) for different groups of speakers (e.g. by speaker sex), as well as in different regions of the parameter range for another variable (e.g. tongue area). We also comment on the predictive capacity of these models, showing how they perform when only a subset of the data is used for training and a test set is used for assessing the accuracy of predictions. We conclude by discussing the implications of vocal tract variability for our understanding of evolutionary patterns in human speech.

References.

Fant, Gunnar (1966). "A note on vocal tract size factors and non-uniform F-pattern scalings". In: *Speech Transmission Laboratory Quarterly Progress and Status Report* 1, pp. 22–30.

Fitch, W. Tecumseh and Jay Giedd (1999). "Morphology and development of the human vocal tract: a study using magnetic resonance imaging". In: *Journal of the Acoustical Society of America* 106.3, pp. 1512–1522.

Lammert, Adam, Michael I. Proctor, and Shrikanth S. Narayanan (2013). "Morphological variation in the adult hard palate and posterior pharyngeal wall". In: *Journal of Speech, Language, and Hearing Research* 56.2, pp. 521–530.

Lim, Yongman et al. (2021). "A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images". In: *Scientific Data* 8.187, pp. 1–14.

Schneider, Caroline A., Wayne S. Rasband, and Kevin W. Eliceiri (2012). "NIH Image to ImageJ: 25 years of image analysis". In: *Nature Methods* 9, pp. 671–675.

Stone, Maureen, Jonghye Woo, Junghoon Lee, Tera Poole, Amy Seagraves, Michael Chung, Eric Kim, Emi Z. Murano, Jerry L. Prince, and Silvia S. Blemker (2018). "Structure and variability in human tongue muscle anatomy". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.5, pp. 499–507.

Production and perception of tonal coarticulation: Evidence from computational simulation of communication

Huang, Po-Hsuan

Graduate Institute of Linguistics, National Taiwan University

benson32169@gmail.com

Introduction. Lexical tones have been found to coarticulate with the preceding and following tones (e.g., Xu 1994; Wang 2002). For instance, Xu has found that tones after high tone offsets (i.e., 55 and 35 tones) were raised, and those after low tone offsets (i.e., 51 and 21 tones) were lowered in Beijing Mandarin (BM). Such tonal coarticulation (TC) has also been found in languages including Taiwan Southern Min (TSM; e.g., Wang 2002), Taiwan Mandarin (TM; e.g., Huang 2023), etc. Variations induced by such coarticulation have also been found to affect listeners' perceptions and cause a target tone to be perceived as other lexical tones (Xu 1994; Wang 2002). This therefore leads to the question of how tone language speakers keep faithful perceptions of tones under TC. In Xu (1994) and Zhang et al. (2022), BM speakers have been found to cope with such variations with normalization, where tones after high offsets were perceived as lower, and those after low offsets as higher. While similar normalizing effects have been found in other languages including TSM and TM (Wang 2002; Huang 2023), it may not be the only mechanism used to cope with tone variations induced by TC. Specifically, it has been proposed by Huang (2023) that linguistic differences may exist with regard to three aspects: 1) the magnitude of TC, 2) the magnitude of normalization for TC, and 3) the ranges of tone acceptance. While Mandarin may allow for TC with the listeners being able to retrieve the target tones back through normalization, it might be less viable for languages with a larger tone inventory, such as TSM. In TSM, the recoverability of the target tones through normalization is lower due to the multiple possible tones that may surface as the same coarticulated tones in the same position. TSM users might alternatively reduce such variations by avoiding the same magnitude of TC as Mandarin, or by maintaining narrower tone acceptance ranges to keep out coarticulated tones. In Huang, the latter was found. While both TM and TSM had similar degrees of TC, TM demonstrated stronger normalization for TC than TSM. On the flip side, TSM maintained narrower tone acceptance ranges as compared with TM. It is argued by Huang that such linguistic differences resulted from the different tone distributions of TM and TSM. However, since the experiments were behavioral experiments conducted on human subjects, multiple aspects need to be factored in, and a direct relation between the linguistic differences and tone distributions could not be easily drawn. Computational simulation of real-world communication may shed light on such an issue. Past studies have proved the ability of communication simulation to capture important linguistic features through the interaction of the speaker and listener agents. In Ren et al. (2020), compositionality emerged through the training of two neural agents simulating the speaker and the listener. Likewise, in Carlsson et al. (2023), it has been found that the joint combination of communication simulation and iterated learning could result in efficient color naming systems similar to those found in human languages. In this study, the author uses a speaker neural agent and a listener neural agent, modulated by the three aspects proposed by Huang (2023), to simulate real-world tone communication under TC, and seeks to provide a more direct observation of the relation between tone distributions and the linguistic differences in the production and perception of tonal coarticulation.

Methods. To simulate tone communication, tone contours were represented as tone onsets and offsets ranging from 1-5, based on the five-level tone marks. In TM, there were four tone contours: (5, 5), (3, 5), (2, 1), and (5, 1). In TSM, an additional (3, 3) was added, leading to a larger inventory. The data were generated based on these tones as tone contours, and a TM model and a TSM model were trained. To simulate real-world variance, for each tone contour generation, the tone onset and offset were randomly sampled on normal distributions with the means being the standard values. To simulate TC between the preceding and target tones, for each possible combination, 2048 tokens (tone contour pairs) were generated. Among them, 80% were taken as the training set, and 20% were taken as the test set. To train the listener agent to recognize the canonical tone contours (i.e., not affected by TC), additional 2048 tokens (singleton tone contours) were generated for each tone. Two neural agents were constructed with multilayer perceptrons (MLP) and trainable parameters to represent the speaker and the listener in communication. During each epoch, the listener was first trained with singleton tone contours for four sub-epochs (Phase A) to recognize the canonical contours of the tones, and then joined with the speaker and trained with tone contour pairs for another four sub-epochs (Phase B). There were a total of 256 epochs. In the speaker agent, an MLP was used to simulate coarticulation. A tone contour pair was taken as the input, and then a value from 0-1 was produced by the MLP and used as the degree of coarticulation with which the target tone would be coarticulated. A value of 1 meant complete coarticulation with the preceding tone, while 0 meant no coarticulation at all. The coarticulated tone contour pairs would then be taken as the input for the listener (in Phase B). In the listener agent, two trainable parameters and two MLPs were used. The trainable parameters represented the tone acceptance ranges as normal distributions for each lexical tone, with one being the mean value of the acceptance range, and the other being the standard deviation. For the two MLPs, one was used to represent phonological perception. The

contours of the target tone were first converted into tone acceptances (the probability for this contour to be accepted as each of the lexical tone based on the acceptance distributions) and then taken by the phonological perception MLP, and an initial prediction of which tones this token might be was produced. If the training was in Phase A, this initial prediction would be directly used as the final prediction and evaluated for backpropagation. If it was in Phase B, the initial prediction would be joined with the contours of the preceding tone as input for the other MLP, which was used to allow for the normalization of the listener and the final prediction of the target tone would be produced. Lastly, the three aspects proposed by Huang (2023) were evaluated. The mean degree of coarticulation of the speaker for the test set was taken as the magnitude of coarticulation. For the normalization for TC, following Zhang et al. (2022), a series of target tones simulating a continuum from the low tone (21) to the falling tone (51) (i.e., with the onset of the target tone going from low to high and the offset staying at 1) following different preceding tones were predicted by the TM and TSM models. If normalization was at work, a preceding tone with high offsets (e.g., 55 and 35) would lead the target tone to be perceived as lower, and it would have to be very close to a canonical 51 to be perceived falling, and vice versa. The stronger the normalization, the larger this effect would be. Finally, tone acceptance ranges were assessed by the standard deviations of the tone acceptance ranges during validation. A smaller deviation would indicate a narrower tone acceptance range.

Results. The accuracies of the TM model and TSM models were 0.62 and 0.54. This relatively low performance was understandable since variances were intentionally introduced to mimic real-world speech. The mean degrees of coarticulation in the TM and TSM models were 0.43 and 0.46. TSM therefore did not seem to have a lower magnitude of coarticulation. The magnitudes of normalization for TC in the two models are shown in **Figure 1**. As can be seen, compared with TSM, the TM model was more subject to the offset height of the preceding tones, as indicated by the interval (1.40) between the orange line (51 as the preceding tone) and the green/blue lines (35 and 55 as the preceding tones) at the 0.5 midpoint, which is much larger than the one in the TSM model (1.20). Finally, the mean standard deviations of the TM and TSM models were 2.30 and 2.28, suggesting generally narrower tone acceptance ranges in the TSM model than in the TM model.

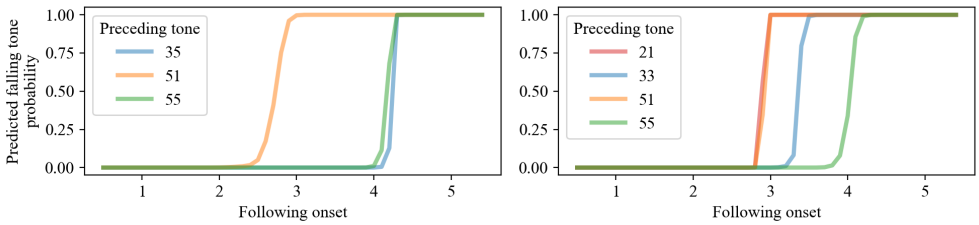


Figure 1:*Normalization of the listener neural agent for different preceding tones on a low-to-falling tone continuum (left: TM; right: TSM).*

Discussion. The results of the simulation in general supported the hypothesis that different tone distributions could lead to different strategies in dealing with the tone variations induced by TC. Specifically, by simulating the respective tone inventories of TM and TSM, the two models largely replicated the findings in Huang (2023) on human subjects. Similar to Huang's production experiment results, the degrees of TC in the two models were rather comparable; while TSM had a more complex tone inventory, it did not lead to a smaller degree of TC in TSM. On the flip side, linguistic differences were found in terms of perception. Like TSM speakers, the TSM model demonstrated a smaller magnitude of normalization for TC, as compared with the TM model. Similarly, like its real-world counterparts, the TSM model maintained generally narrower ranges of tone acceptance than the TM model. In Huang (2023), it is explained that, while TSM could not rely as much on normalization as Mandarin due to the lower recoverability of the target tones, it could use stricter tone acceptance ranges to filter out coarticulated tones that could potentially be confused with other lexical tones, in turn reducing its reliance on normalization. The results of the simulation in this study therefore supported such an explanation by manipulating the tone distributions of the models. In general, this study demonstrates the possibility of simulating real-world communication and human cognitive mechanisms as well as its ability to allow for more direct explanations of production and perception behaviors through the interaction of the speaker and listener agents.

References

Carlsson, E., Dubhashi, D.P., & Regier, T. (2023). Iterated learning and communication jointly explain efficient color naming systems. *ArXiv*, abs/2305.10154.

Huang, P.H. (2023). Perception and Production of Coarticulated Tones in Taiwan Mandarin and Taiwan Southern Min [Master's thesis]. National Taiwan University, Taipei.

Ren, Y., Guo, S., Labeau, M., Cohen S.B., & Kirby, S. (2020). Compositional languages emerge in a neural iterated learning model. International Conference on Learning Representations, online.

Wang, H. (2002). The prosodic effects on Taiwan Min tones. *Language and Linguistics*, *3*, 839-852.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America*, *95*(4), 2240-2253.

Zhang, H., Ding, H., & Lee, W.-S. (2022). The influence of preceding speech and non-speech contexts on Mandarin tone identification. *Journal of Phonetics*, *93*, 101154

Inter-gestural coupling of onset and vowel gestures in adults who stutter in different rhythmic conditions

Mona Franke[1,2,3], Simone Falk[2,3], Philip Hoole[1]

[1]Institute for Phonetics and Speech Processing, LMU München, Germany,
[2]Faculté des arts et des sciences – Départment de linguistique et de traduction, UdeM, Montréal, Canada,
[3]International Laboratory for Brain, Music, and Sound Research (BRAMS), Montréal, Canada
Mona.franke@phonetik.uni-muenchen.de

Introduction. This study investigates inter-gestural timing, focusing on the coupling between onset (consonant) and vowel gestures in persons who stutter (PWS) and persons who do not stutter (PWNS). Grounded in Articulatory Phonology, our research explores potential disruptions in speech motor control associated with stuttering, a neurodevelopmental speech motor disorder linked to differences in the planning and execution of speech movements (Smith & Weber 2016). These differences lead to speech disfluencies like repetitions, prolongations, and blocks of sounds or syllables, occurring predominantly at the onset of stressed words or syllables and maximally reach the onset of the vowel (Harrington 1987). Additionally, perceptually fluent speech of PWS is temporally more variable than that of PWNS. Previous studies identified various articulatory differences in PWS, including increased variability in lip-jaw coordination, decreased peak velocity from onset to the following vowel, higher inter-articulator variability, and longer movement durations of lip- and jaw-closing movements (De Nil 1995; Heyde et al. 2016; Smith et al. 2010; Max et al. 2003). Consonant-vowel coordination is especially interesting to investigate in PWS (Harrington 1987, 1988) but also in PWNS as the coordination between onset consonant gestures and vowel gestures is more cohesive compared to the coordination between vowel gestures and coda consonant gestures (see Hoole & Pouplier 2015, for review).
There is supporting evidence indicating that the timing between two distinct articulatory movements, i.e. inter-gestural timing, can be influenced by an external rhythm (Tilsen 2009). One fascinating phenomenon in this respect is that PWS often demonstrate increased speech fluency when speaking alongside an external rhythm, like a metronome (Wingate 1969). Nevertheless, the articulatory basis of this phenomenon remains unclear. To address this gap, our study investigates inter-gestural coupling in fluent speech of PWS compared to PWNS and explores the influence of different rhythmic conditions—motor pacing, auditory pacing, and a combination of both on consonant-vowel (CV) coupling.
We hypothesize that PWS show differences in inter-gestural timing, potentially attributed to articulatory timing differences. Auditory pacing is expected to mitigate these differences, while motor pacing is anticipated to stabilize overall articulatory timing (Parrell et al. 2014). Hence, we would not expect to find a group difference in the motor pacing condition. On the contrary, the simultaneous use of auditory-motor pacing might lead to increased timing variability in PWS, as previous findings indicate heightened variability in synchronizing speech and hand movements to a tone (Hulstijn et al. 1992). The latter authors suggest that PWS have challenges in response coordination. To our knowledge, since their study no other study has included such a range of conditions. The key feature of our data set is thus the ability to study articulatory processes, particularly the critical phase of CV coordination, in different rhythmic conditions.

Methods. Ten adults who stutter and ten adults who do not stutter participated in this study. All participants were native speakers of German and the groups were age- and sex matched, as well as matched for handedness. Electromagnetic Articulography (EMA) data was collected while participants produced mono- and disyllabic German target words. In total, there were three monosyllabic and three disyllabic target words that all started with a bilabial onset ([m] or [b]), followed by a back vowel ([o:] or [u:]), embedded in the carrier phrase [ˈzeːə ____ ˈan] (Look at ____).
The experiment included four conditions, each involving the repetition of target words four times within a quasi-randomized order, along with filler words. The conditions were conducted in the following order:

- Baseline: Reading words within a self-chosen speech tempo.
- Tapping (self-paced): Baseline condition with the added task of aligning a finger tap to each word.
- Metronome (externally paced): Reading words synchronized to a metronome (90 bpm).
- Metronome+Tapping: Reading words while tapping a finger to each word and synchronizing to a metronome (90bpm).

In tapping conditions, participants were instructed to tap their dominant hand's index finger on an elevated wooden block placed on a nearby table. Sensors relevant for the reported data were positioned on the tip of the index finger, upper and lower lip, as well as the tongue back (TBACK), and tongue mid (TMID).
The constriction for the bilabial onset was measured using Lip Aperture (LipAp). A well-defined reference point for the bilabial constriction is the velocity maximum related to the closing movement, which was semi-automatically detected. The target of the vowel gesture was segmented based on the horizontal movement of the TBACK sensor by detecting the maximum constriction. Additionally, an anchor point was set in the /e:/ of the carrier word /zeːə/ by detecting the tangential velocity minimum of the TMID sensor.

In order to measure inter-gestural timing, we have started with a purely spatial measure, as the articulatory vowel onset is quite challenging to detect (note that typical lag measures are in preparation). The spatial measure is based on the amount of progress from the anchor point towards the TBACK target at the time of the bilabial reference point. The relative distance from the position of TBACK at the time point of the bilabial constriction relative to /e:/ in the carrier phrase (D1) and the target vowel /u:/ and /o:/ (D2) was calculated for each target word per condition ($\log_{10}$(D1/D2)). There is an illustration of this measure in the left panel of **Figure 1**. A distance of 0 would indicate that the TBACK sensor at the time point of the bilabial constriction has the same distance to the anchor point and to the TBACK target (vowel). The more negative the value, the greater the distance to the target vowel still remaining at the time of the bilabial constriction.

Results. To date, two participants per group have been analyzed; we expect to have results for the entire group by spring 2024. **Figure 1** displays an example of the spatial analysis for the target vowel /u:/ of one participant in the Metronome+Tapping condition (left panel) and results for all target words for one participant pair (right panel).

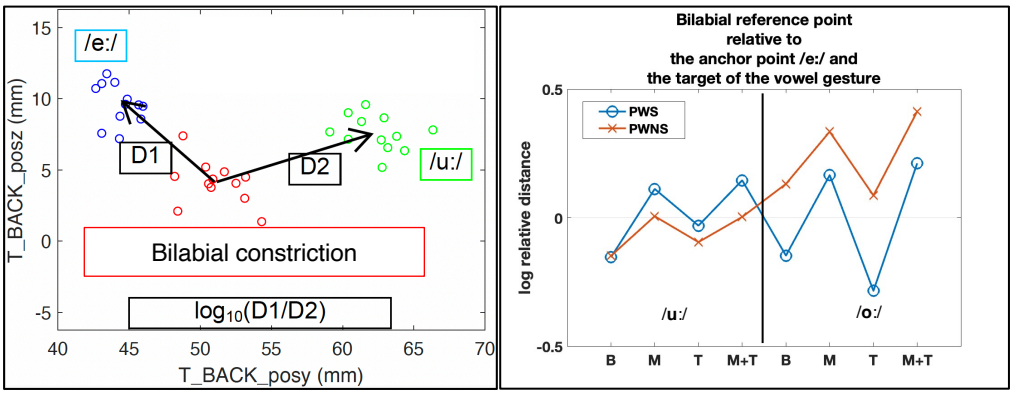


Figure 1: *Left panel: Illustration for spatial measure. Right panel: Mean of relative distance for all target words for one participant pair in different conditions (B=Baseline, M=Metronome, T=Tapping, M+T=Metronome+Tapping).*

In both the Baseline and the Tapping condition, there is less progress towards the following vowel at the time of the bilabial constriction compared to the conditions with a metronome. This is consistent across all participants and vowels. With only two participant pairs analyzed, conclusions about potential differences between PWS and PWNS are premature.

Discussion. It is unclear from previous work whether to expect longer (e.g. Verdurand et al., 2020) or shorter (e.g. Harrington, 1988) onset-vowel lags in PWS. We will discuss our findings in light of theories of stuttering and fluent speech production. In addition, we will discuss a possible modulation of CV coordination in the context of a pacing condition. Our study contributes to the evolving understanding of the fine interplay between consonant and vowel gestures in fluent speech production, with a specific focus on how these dynamics may differ between PWS and PWNS.

References

De Nil, L. F. (1995). The influence of phonetic context on temporal sequencing of upper lip, lower lip, and jaw peak velocity and movement onset during bilabial consonants in stuttering and nonstuttering adults. *Journal of Fluency Disorders*, 2, 127-144.

Harrington, J. M. (1987). Coarticulation and stuttering: an acoustic and electropalatographic study. In H. Peters, & W. Hulstijn (eds.), *Speech motor dynamics in stuttering*. New York: Springer Verlag.

Harrington, J. (1988). Stuttering, Delayed Auditory Feedback, and Linguistic Rhythm. *Journal of Speech and Hearing Research*, 31, 36-47.

Heyde, C. J., Scobbie, J. M., Lickley, R. & Drake, E. K. E. (2016). How fluent is the fluent speech of people who stutter? A new approach to measuring kinematics with ultrasound, *Clinical Linguistics & Phonetics*, 30:3-5, 292-312

Hoole, P., & Pouplier, M. (2015). *Interarticulatory coordination - speech sounds*. In M. Redford (Ed.), *The Handbook of Speech Production*, (ch. 7, pp. 133-157). Hoboken, New Jersey: John Wiley & Sons.

Hulstijn, W., Summers, J. J., van Lieshout, P. H. M, Peters, H. F. M. 1992. Timing in finger tapping and speech: A comparison between stutterers and fluent speakers. *Human Movement Science*, 11(1–2), 113–124.

Max, L., Caruso, A. J., & Gracco, V. L. (2003). Kinematic analyses of speech, orofacial nonspeech, and finger movements in stuttering and nonstuttering individuals. *Journal of Speech, Language, and Hearing Research*, 46, 215-232.

Parrell, B., Goldstein, L., Lee, S., Byrd, D. 2014. Spatiotemporal coupling between speech and manual motor actions. *Journal of phonetics*, 42, 1–11.

Smith, A., Sadagopan, N., Walsh, B., & Weber-Fox, C. (2010). Increasing phonological complexity reveals heightened instability in inter-articulatory coordination in adults who stutter. *Journal of Fluency Disorders*, 35, 1-18.

Smith, A. & Weber, C. (2016). Childhood Stuttering: Where Are We and Where Are We Going? *Semin Speech Lang.,* 37(4), 291-297.

Tilsen, S. (2009). Multitimescale Dynamical Interactions Between Speech Rhythm and Gesture. *Cognitive Science,* 33, 839-879.

Verdurand, M., Rossato, S., & Zmarich, C. (2020). Coarticulatory Aspects of the Fluent Speech of French and Italian People Who Stutter Under Altered Auditory Feedback. *Frontiers in psychology*, *11*, 1745.

Wingate, M. E. (1988). *The Structure of Stuttering (a Psycholinguistic Analysis)*. New-York, NY: Springer Verlag.

How Do Speakers Respond to Altered Formant Feedback Simulating a Change in Speaker Gender?

Erin Doty[1], Douglas Shiller[2], Vesna Novak[3], Tara McAllister[1]

[1]New York University
[2]Université de Montréal
[3]University of Cincinnati

emd14@nyu.edu, douglas.shiller@umontreal.ca, novakdn@ucmail.uc.edu, tkm214@nyu.edu

Introduction. Since Houde and Jordan's (1998) seminal study, auditory-motor adaptation experiments have made important contributions to understanding the relative importance of feedback and feedforward control in speech production. In such studies, the acoustic properties of a talker's voice are altered and played back in near real-time as they read words or word strings. On average, participants alter the acoustic properties of their speech in a direction opposing the shift (sometimes termed compensation), usually without conscious awareness. Changes typically persist for a short time after the altered feedback is withdrawn, which constitutes evidence of adaptation of the speech-motor plan. Previous estimates suggest that the average magnitude of compensation is on the order of 25% of the level of the applied perturbation (Miller et al., 2023), although some people fail to compensate or shift in the same direction as the altered feedback.

A number of studies have used altered auditory feedback to investigate the nature of feedback control and feedforward updating in different clinical populations, such as people who stutter (e.g., Daliri *et al.* 2018) or speakers with dysarthria (e.g., Mollaei *et al.* 2016). Because most speakers produce a robust compensatory response when exposed to altered auditory feedback, it is tempting to try to incorporate altered feedback into clinical efforts to change the motor plans used in speech. In practice, though, such applications have been elusive. This is partly because of the short-term nature of changes induced by altered auditory feedback, but also because the typical type of compensation (e.g., a systematic change in formant frequencies) does not constitute a clinical goal for most populations. This project will explore the idea that altered auditory feedback yielding a systematic change in vowel formants could have clinical relevance in the context of gender-affirming voice training (GAVT).

Transgender and gender-diverse people (henceforth, "trans people") can be negatively impacted if their voice is perceived as incongruous with their gender identity, and they may choose to work independently or with a speech pathologist to achieve a vocal presentation that is comfortable for them. When ascribing gender to a voice, listeners use multiple acoustic cues. The most obvious of these is fundamental frequency (*fo*), which dictates perceived vocal pitch. However, resonant frequencies of the vocal tract (formants) also play an important role in vocal presentation of gender. Women's vocal tracts tend to be shorter than men's, resulting in formants that are roughly 20% higher than men's on average (Childers & Wu 1991). In the GAVT context, the goal is to shift all formants in the same direction, basically mimicking the effects of a smaller or larger vocal tract. However, people find resonance challenging to understand conceptually and to modify in practice (Bush *et al*., 2022). This project will lay a framework for understanding whether altered auditory feedback could serve as a learning tool to support resonance modification in the GAVT context. As our first step, we aim to determine to what extent people compensate for altered auditory feedback in which the first and second formants (F1 and F2) are shifted by a substantial magnitude across all vowels, in a manner consistent with a difference in speaker gender.

Methods. In this study, participants read word strings aloud while auditory perturbation is applied with the Matlab-based program Audapter (Cai *et al.*, 2008). F1 is shifted down by a factor of 1.18 and F2 down by 1.1, scale factors chosen by aggregating across multiple studies comparing vowel spaces of men and women (Vorperian & Kent 2007). The stimuli are six bVd syllables (*bid, bad, bed, bod, booed,* and *bud*), randomly ordered into 5-syllable word strings. Experimental stimuli are elicited in five blocks as follows: baseline (no alteration), ramp up to 90% alteration, hold at 90%, hold at 90%, and washout (no alteration). Each block elicits 50 word-strings, except the final block, which elicits 25. Compensation in each altered feedback block and aftereffect in the washout block are measured acoustically. Our outcome of interest is the component of production change relative to baseline that directly opposes the F1 and F2 shifts. Once data collection is complete, we will fit a mixed-effects linear model with *magnitude of compensation* as the outcome variable and *block* as predictor variable, with random effects of *participant* and *word string*.

The participants in this study will be 20 cisgender men aged 18-65. Cis men rather than trans women will be recruited for this initial study due to the challenges of recruiting from a small and minoritized population. We will use broad inclusionary criteria in order to maintain comparability with transgender participants to be recruited in future studies. Language background criteria will be flexible, as long as participants are proficient in English, based on previous evidence

that language background does not impact response to altered feedback (Shiller *et al.* 2023). Participants will be screened for a history of speech/language disorder, as well for significant discomfort reading out loud. Participants must also pass a hearing screening at 25 dB.

Results. To date, our task has been piloted with four cis men. Results showed that on average, participants compensated and adapted by shifting their produced F1 and F2 in the opposite direction of the applied auditory perturbation by up to roughly 25% of the applied perturbation. While this resulting change is small relative to the average difference in formants between men and women, even small-magnitude formant shifts have been reported to impact listener attribution of gender (Gallena *et al.* 2018). **Figure 1** shows the response of one pilot participant to auditory perturbation of F1 and F2. The black arrows indicate the applied perturbation, and the blue arrows show the subject's compensatory shift of F1 and F2 during each trial. The red arrows show the average of those responses over the whole block.

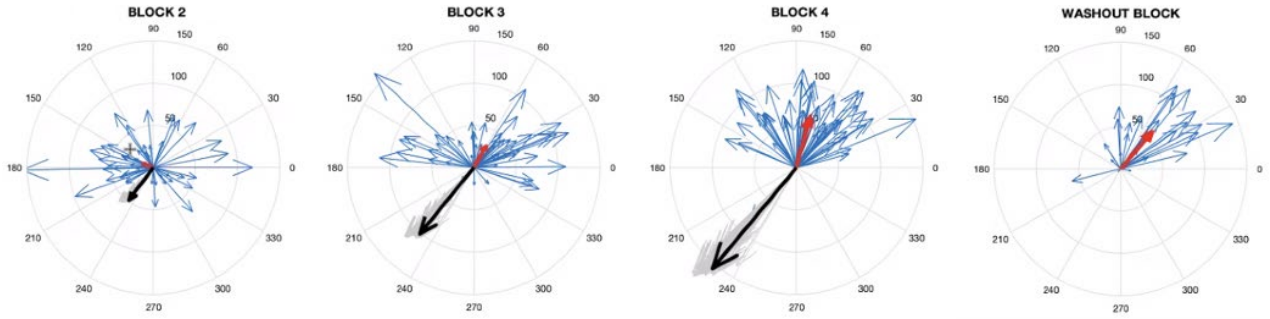


Figure 1: One pilot participant's response to a perturbation in which F1 and F2 were shifted in the same direction with a large magnitude, simulating a change in effective vocal tract length.

Discussion. We have successfully piloted the experiment and will collect data from the proposed 20 participants in January 2024. Based on our pilot data, we hypothesize that participants will exhibit a significant degree of compensation to the perturbation applied to F1 and F2, and that some degree of shift will persist after the perturbation is withdrawn. While we will begin by using change in F1 and F2 to index response to altered feedback, our next step will be to collect blinded listeners' ratings to determine whether any acoustic changes are perceived as impacting the speakers' gender presentation (e.g., more masculine, more feminine). If our research suggests that adaptation to altered feedback could help learners identify motor plans to achieve an altered vowel space, follow-up studies will investigate whether learners could be trained to generalize these motor plans beyond the immediate context of altered auditory feedback.

References

Bush, E. J., Krueger, B. I., Cody, M., Clapp, J. D., & Novak, V. D. (2022). Considerations for voice and communication training software for transgender and nonbinary people. *Journal of Voice*. Early online: doi:https://doi.org/10.1016/j.jvoice.2022.03.002

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2008). A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iau/. Proceedings of the 8th International Seminar on Speech Production, 65-68.

Childers, D. G., & Wu, K. (1991). Gender recognition from speech. Part II: Fine analysis. *Journal of the Acoustical Society of America, 90*(4 Pt 1), 1841-1856. doi:10.1121/1.401664

Gallena, S. J. K., Stickels, B., & Stickels, E. (2018). Gender perception after raising vowel fundamental and formant frequencies: Considerations for oral resonance research. *Journal of Voice*, *32*(5), 592-601. doi:10.1016/j.jvoice.2017.06.023

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, 279(5354), 1213-1216. doi: 10.1126/science. 279.5354.1213

Miller, H. E., Kearney, E., Nieto-Castañón, A., Falsini, R., Abur, D., Acosta, A., ... & Guenther, F. H. (2023). Do not cut off your tail: a mega-analysis of responses to auditory perturbation experiments. J*ournal of Speech, Language, and Hearing Research, 66*(11), 4315-4331.

Mollaei, F., Shiller, D. M., Baum, S. R., & Gracco, V. L. (2016). Sensorimotor control of vocal pitch and formant frequencies in Parkinson's disease. *Brain Research, 1646*, 269-277.

Shiller, D. M., Bobbitt, S., & Lametti, D. R. (2023). Immediate cross-language transfer of novel articulatory plans in bilingual speech. *Journal of Experimental Psychology: General*. Early online: https://psycnet.apa.org/doi/10.1037/xge0001456

Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: a synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research, 50*(6), 1510-1545. doi:10.1044/1092-4388(2007/104)

Control and Recovery of Vocal Tract Gestures Using Stochastic Models of Action and Perception: The Martingale Dynamics of Human Speech

Gordon Ramsay[1]

[1]*Spoken Communication Laboratory, Department of Pediatrics, Emory University, Atlanta, Georgia, USA*
gordon.ramsay@emory.edu

Introduction.
Models of speech production typically assume that goal-directed target trajectories centrally specified through feedforward planning are realized by peripheral servomechanisms via feedback control comparing predictions from an internal model with somatosensory and auditory signals from body and environment. Conversely, models of speech perception are typically based on analysis-by-synthesis, where hypotheses from a generative model predicting the sensory consequences of possible action sequences are evaluated against actual sensory input to recover the most likely phonological or articulatory targets from sound. In both production and perception, the key assumption is that the control and recovery of vocal tract gestures is based on monitoring the *prediction error* measuring the discrepancy between the output of an internal model and available input. In this paper, we show that the structure of both action and perception systems can be derived using martingale calculus from basic principles of stochastic estimation and control, which is indeed based on predictive simulation, but need not depend on explicitly measuring prediction error as previously proposed.

Methods.
Assume an underlying probability space $(\Omega, \mathcal{F}, P)$. Let $X = \{X_t : t \in \mathbb{R}_+\}$ be a stochastic process representing the physical state of the vocal tract and control system, and let $Y = \{Y_t : t \in \mathbb{R}_+\}$ be a stochastic process representing partial somatosensory and auditory measurements of the state, with $\mathcal{F}^X$ and $\mathcal{F}^Y$ the right-continuous filtrations generated by X and Y. Provided that X and Y have finite mean variation, the Doob-Meyer Theorem guarantees that each can be decomposed into *predictable* and *unpredictable* components, and written as stochastic differential equations:

$$dX_t = g(X_t, t)dt + v(t)dV_t, \quad (1)$$

$$dY_t = h(X_t, t)dt + w(t)dW_t, \quad (2)$$

where $V = \{V_t : t \in \mathbb{R}_+\}$ and $W = \{W_t : t \in \mathbb{R}_+\}$ are martingale processes that reflect the nature of random variation in the system, and g, h, v, w are appropriate measurable functions representing the physical dynamics of the state trajectories. Properties of the system are defined as functionals $p_t(\phi) := E\{\phi(X_t)\}$ of the system state X, given by:

$$dp_t(\phi) = p_t(L\phi)dt, \quad (3)$$

where L is the infinitesimal generator of (1), describing how probability mass is transported along the sample paths. Equation (3) is essentially the "forward model" recursively characterising the statistical dynamics of the system, and the unconditional probability law of the state can be recovered as a special case since $P(A) = p_t(I_A)$. When only partial information about the state is available indirectly by observing Y, the optimal estimate of any functional of the state can be shown to be given by the conditional mean $\pi_t(\phi) := E\{\phi(X_t)|\mathcal{F}_t^Y\}$, with the conditional probability law $P(A|\mathcal{F}_t^Y) = \pi_t(I_A)$ as a special case. Two key approaches can be used to characterize the stochastic functional π_t. According to the Fujisaki-Kallianpur-Kunita Theorem, π_t is generated by the stochastic differential equation:

$$d\pi_t(\phi) = \pi_t(L\phi)dt + \sigma_t(h, \phi)d\nu_t, \quad (4)$$

where $\sigma_t(h, \phi) := \pi_t(h\phi) - \pi_t(h)\pi_t(\phi)$ is the *conditional covariance* and $\nu_t = Y_t - \int_0^t \pi_s(h)ds$ is the *innovations process*, the discrepancy between the actual and predicted measurements. Alternatively, according to the Duncan-Mortenson-Zakai Theorem, there exists a probability measure $\overline{P}$ under which the functional $\overline{\pi}_t := \overline{E}\{\phi(X_t)|\mathcal{F}_t^Y\}$ evolves according to:

$$d\overline{\pi}(\phi) = \overline{\pi}_t(L\phi)dt + \overline{\sigma}_t(h, \phi)dY_t, \quad (5)$$

where $\overline{\sigma}_t(h,\phi) := \overline{\pi}_t(h\phi)$ is the *conditional correlation* and $\pi_t(\phi) = \overline{\pi}_t(\phi)/\overline{\pi}_t(1)$. Equations (4) and (5) are mathematically equivalent but have different interpretations. The first term in both equations is the best predictor of the current state given the past measurement history, and can be considered to be a stochastic "internal model" of the vocal tract. Equation (4) implements a *predictor-corrector* structure, where predictions of the internal model are corrected according to measurements of the prediction error (the innovation process ν) between the predicted and actual measurements, weighted by an adaptive gain given by the conditional covariance. Equation (5) implements a *predictor-correlator* structure, where predictions of the internal model are weighted directly by the measurements, according to an adaptive gain given by the conditional correlation. The predictor-corrector structure is consistent with current proposals for speech motor control and analysis-by-synthesis, and relies on explicitly calculating and monitoring a prediction error signal; the Kalman filter analogy proposed by many authors is a special case for linear systems, whereas the result stated here is more general. The predictor-correlator structure is novel and processes sensory input directly without calculating prediction error; the adaptive gain term given by the conditional correlation can be considered as a form of "salience map" that adaptively assigns importance to different regions of the measurement trajectory space according to how these resonate with the underlying dynamical system. Although both of these results are possible as mathematical models of action and perception, the predictor-correlator structure is considerably simpler and more plausible neurophysiologically.

Results.

To provide a concrete illustration of this framework, we revisit a stochastic target trajectory model of speech production and perception we proposed previously for articulatory synthesis and recognition, based on gestural phonology and the equilibrium-point model of speech motor control (Figure 1). In our model, sequences of probabilistically timed gestural symbols are generated by a semi-Markov chain. Each gesture is associated with a probability distribution of target trajectory parameters, which are sampled each time a gestural state is entered to generate a piecewise-deterministic trajectory of control parameters then used to drive a biomechanical plant. Each target distribution in the input space describes the probability that particular articulatory or acoustic correlates that characterize the corresponding gesture will be successfully realized in the output space by particular control trajectories. A biomechanical simulation generates trajectories of articulatory parameters from control trajectories, and an acoustic simulation converts the resulting time-varying vocal tract shapes into sound. The model structure is a special case of equations (1)-(2). It can be used as a model of speech production for articulatory synthesis, by randomly sampling the probability space according to equation (3), and can also be used as a model of speech perception for articulatory speech recognition, by solving equations (4)-(5). The key feature of the model is that it does not assume invariance at any level of the processes underlying the production and perception of speech; instead, the objects of speech production and perception are joint probability distributions on any available representational spaces, conditioned on any available information. Previous implementations were based on approximations of (4); here we present and evaluate numerical simulations of (5) based on particle filtering. Using a finite-state grammar of simple VCV sequences, we show how random sampling of the model states can be used to generate plausible statistical patterns of coarticulatory variability in articulation and acoustics. By calculating the conditional probability distributions of the model states, conditioned on the acoustic trajectories, we show how the time-varying salience maps that arise within the predictor-correlator structure correctly attribute and recover gestural information from sound.

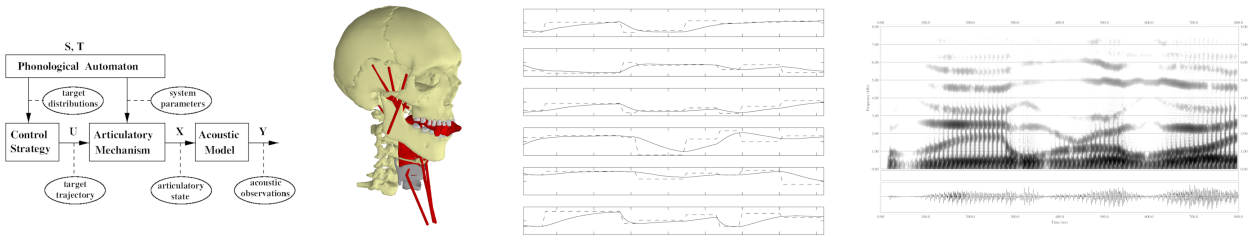


Figure 1: *Stochastic target trajectory model, showing control structures, biomechanical plant, and acoustic synthesis.*

Discussion.

We provided an overview of optimal estimation for stochastic dynamical systems, based on martingale calculus and nonlinear filtering, and we showed that the results of this theory, derived from first principles entirely from mathematics, are consistent with theories of speech production and perception based on predictor-corrector analogies, derived from empirical evidence. However, we also showed that an alternative and simpler account is also possible, based on a predictor-correlator structure, where predictive simulations based on internal models adaptively weight sensory input directly.

Spatiotemporal features of bilabial geminate and singleton consonants in Italian

Francesco Burroni[1], Sireemas Maspong[1], Nicole Benker[1], Phil Hoole[1], James Kirby[1]

[1]Institute for Phonetics and Speech Processing, LMU Munich, Germany

{francesco.burroni|s.maspong|hoole|jkirby}@phonetik.uni-muenchen.de
nicole.benker@campus.lmu.de

Introduction. Italian speakers employ consonantal duration contrastively, e.g., It. [papa] "pope" vs. It. [papːa] "(baby) food". A substantial body of work has investigated the acoustic correlates of such singleton *vs*. geminate contrasts (Di Benedetto et al., 2021 for a review); much less work has been dedicated to their kinematic underpinnings (Celata et al., 2022 for a review). In addition, previous kinematic work on geminates in Italian has yielded conflicting results (Celata et al., 2022). Beyond robust differences in movements' duration (Zmarich et al., 2011) – which are in line with claims that geminates differ from singletons purely in longer gestural activation intervals (Goldstein, n.d.; Tilsen, 2016) –, it remains unclear whether / how articulatory movements underlying the production of singleton and geminate consonants may differ in terms of both their spatial and temporal features. On the spatial side, it is not clear whether geminates are produced with different articulator movements' target, amplitude, velocity, and stiffness compared to singletons (Celata et al., 2022; Dunn, 1993; Gili-Fivela et al., 2007; Hagedorn et al., 2011; Lofqvist, 2007; Smith, 1992; Zmarich et al., 2011). On the temporal side, it remains to be ascertained whether the timing organization of geminates to surrounding vowels may also be different from singletons (Celata et al., 2022; Smith, 1992; Tilsen & Hermes, 2020; Zmarich et al., 2011). Furthermore, it has also been suggested that differences between singleton and geminates may emerge or disappear under rate manipulations (Tilsen & Hermes, 2020; Zmarich et al., 2011). We present the results of an electromagnetic articulography (EMA) investigation of Italian geminates produced under rate manipulation. We show that geminates are produced with distinct spatial features and a distinct timing regime to surrounding vowels. **Methods.** We collected simultaneous audio and EMA (3D Carsten AG501) data from 10 native Italian speakers, speaking Central and Southern varieties (south of the Rimini-La Spezia line) where geminates are uncontroversially realized (e.g., Mairano & De Iacovo, 2020). Participants produced six disyllabic nonce VCV words containing all singleton and geminate Italian bilabial consonants: [ipa, ipːa, iba, ibːa, ima, imːa]. We refer to /i/ as V_1 and /a/ as V_2. A high to low vowel transition was chosen to maximize tongue vertical movement and facilitate landmarking. Bilabial consonants were chosen so as to avoid competing demands on tongue movement from consonants and vowels. Target words were embedded in a carrier sentence [dika __ due vɔlte] "Please say __ twice". Trials were produced at 5 rates "very slow", "slow", "normal", "fast", "very fast". Each word was repeated 12 times at each rate. We thus collected ~360 tokens per speaker. Bilabial consonants' closure (CLO) and release (REL) phases were identified using a lip aperture (LA) time series. LA was defined as the 3D Euclidean distance between the Lower Lip and Upper Lip sensors. Vocalic gestures were identified on the basis of the first principal component of tongue movement (e.g., as in Sorensen & Gafos, 2015), obtained by entering three dimension movement components of the Tongue Tip, Tongue Body, and Tongue Back sensors in a principal component analysis. Bilabial closure and release were landmarked using velocity zero-crossings to avoid biases in estimating onsets in the presence of intrinsic differences in stiffness (Fig. 1). Vocalic gesture landmarks were identified using a 20% threshold on peak velocity (Fig. 1). From landmarking we extracted 11 "spatial" variables and 5 "lag" variables: (1) Maximum constriction degree of LA; (2-3) Duration of the consonantal closure and release gestures; (4-5) Movement amplitude of closure and release; (6-7) Peak velocity of closure and release; (8-9) Stiffness of closure and release (ratio of peak velocity divided by movement amplitude); (10-11) Time to peak velocity of closure and release; (12) V_1-V_2 lag; (13) V_1-CLO lag; (14) V_1-REL lag; (15) CLO-V_2 lag; (16) REL- V_2 lag. All dependent variables were entered in linear mixed effect regression models. The fixed effects were utterance duration (continuous, z-scored and from which the target segment duration had been subtracted, following the recommendations of Tilsen & Tiede, 2023), geminate status (categorical, with reference as singleton), and their interaction. Maximal random effect structures (with both intercepts and slopes) for subject session and voicing/manner, i.e., whether the consonant is [p], [b], or [m], were also included. **Results.** Concerning spatial features, we found that geminates are produced with more constricted targets (Fig. 2), longer closure and release phases, faster peak velocities in the closure and release phases, wider movement amplitudes in the closure and release phases, lower stiffness in the closure and release phases, and longer times to peak velocities in the closure and release phases (Table 1). Additionally, such properties are differently modulated by rate for singleton vs. geminate consonants. Concerning temporal features, we found that the lag between V_1-V_2, V_1-CLO, V_1-REL are shorter for geminates vs. singletons, while the CLO-V_2 and REL- V_2 lags are longer for geminates than singletons (Fig. 3). **Discussion.** Singleton and geminate bilabial consonants in Italian differ (i) in terms of the kinematic parameters determining their spatial features and (ii) their timing to surrounding vowels. Distinct interactions with speech rate for singleton vs. geminates suggest the differences are in intrinsic parameters. Given these findings, characterizations of Italian geminates in terms of longer gestural activations seem inadequate. Thus, we present trajectory modeling relying on differences in both dynamical parameters, such as higher targets and lower stiffness (Löfqvist, 2005) and also in timing control strategies (Tilsen & Hermes, 2020).

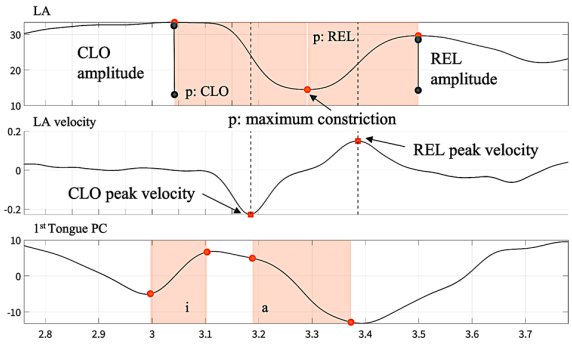


Fig. 1 Example of individual token landmarking for [ip:a].

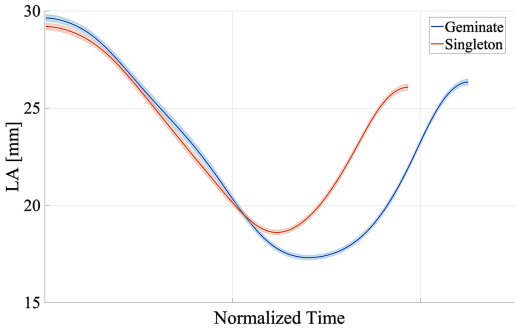


Fig. 2 Average (separately) time-normalized Lip Aperture (LA) trajectories for [iC:a] and [iCa].

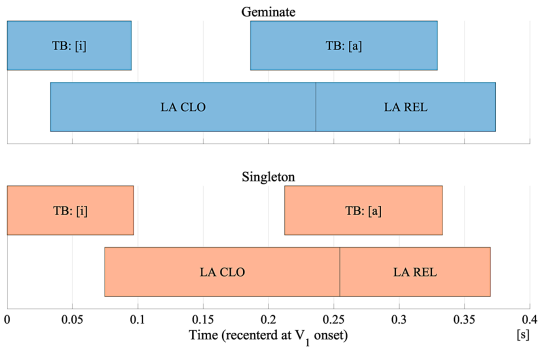


Fig. 3 Gestural scores based on mean gesture duration and mean intergestural lag duration for [iC:a] and [iCa]. Time is recentered to start at V1 onset.

Table 1: *Summary of findings*

	Closure	Release
Target	G > S	
Duration	G > S	G > S
Peak velocity	G > S	G > S
Movement amplitude	G > S	G > S
Stiffness	G < S	G < S
Time to peak velocity	G > S	G > S

References

Celata, C., Meluzzi, C., & Bertini, C. (2022). Acoustic and kinematic correlates of heterosyllabicity in different phonological contexts. *Language and Speech*, *65*(3), 755–780.

Di Benedetto, M.-G., Shattuck-Hufnagel, S., De Nardis, L., Budoni, S., Arango, J., Chan, I., & DeCaprio, A. (2021). Lexical and syntactic gemination in Italian consonants—Does a geminate Italian consonant consist of a repeated or a strengthened consonant? *The Journal of the Acoustical Society of America*, *149*(5), 3375–3386. https://doi.org/10.1121/10.0004987

Dunn, M. H. (1993). *The phonetics and phonology of geminate consonants: A production study*. Yale University.

Gili-Fivela, B., Zmarich, C., Perrier, P., Savariaux, C., & Tisato, G. (2007). Acoustic and kinematic correlates of phonological length contrast in Italian consonants. *ICPhS 2007-16th International Congress of Phonetic Sciences*, 469–472.

Goldstein, L. M. (n.d.). *Dynamical parameters and geminates*.

Hagedorn, C., Proctor, M. I., & Goldstein, L. (2011). Automatic Analysis of Singleton and Geminate Consonant Articulation Using Real-Time Magnetic Resonance Imaging. *INTERSPEECH*, 409–412.

Löfqvist, A. (2005). Lip kinematics in long and short stop and fricative consonants. *The Journal of the Acoustical Society of America*, *117*(2), 858–878.

Löfqvist, A. (2007). Italian geminates: Interarticulator programming and articulatory dynamics. *The Journal of the Acoustical Society of America*, *122*(5_Supplement), 2994–2995.

Mairano, P., & De Iacovo, V. (2020). Gemination in northern versus central and southern varieties of Italian: A corpus-based investigation. *Language and Speech*, *63*(3), 608–634.

Smith, C. L. (1992). *The timing of vowel and consonant gestures*. Yale University.

Sorensen, T., & Gafos, A. I. (2015). Changes in vowel velocity profile with vowel-consonant overlap. *ICPhS*.

Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics*, *55*, 53–77.

Tilsen, S., & Hermes, A. (2020). Nonlinear effects of speech rate on articulatory timing in singletons and geminates. *12th International Seminar on Speech Production*.

Tilsen, S., & Tiede, M. (2023). Parameters of unit-based measures of speech rate. *Speech Communication*, *150*, 73–97. https://doi.org/10.1016/j.specom.2023.05.006

Zmarich, C., Gili-Fivela, B., Perrier, P., Savariaux, C., & Tisato, G. (2011). Speech timing organization for the phonological length contrast in Italian consonants. *Interspeech 2011-12th Annual Conference of the International Speech Communication Association*, 401–404.

Are frequency effects cumulative from word to syllable?

Ivan Yuen, Bistra Andreeva, Omnia Ibrahim, Bernd Möbius

Saarland University

ivyuen@lst.uni-saarland.de, andreeva@lst.uni-saarlande.de, omnia@lst.uni-saarland.de, moebius@lst.uni-saarland.de

Introduction. Although predictability effects have been shown to affect acoustic realization (e.g., Arnon & Cohen-Priva, 2013; Aylett & Turk, 2004) at different linguistic levels, they are mostly based on manipulation of predictability at a specific linguistic level, rather than across levels. 'Frequency of occurrence' is one type of predictability, which could occur at the level of word (word frequency) and syllable (syllable frequency). These frequency effects are assumed to arise from the ease of retrieval, leading to fast response latency (RT). It is postulated that high frequently-occurring syllables are stored in a mental syllabary that mediates phonological and phonetic encoding (e.g., Cholin et al., 2006). Since a word can be mono- or poly-syllabic, this raises questions about the loci of frequency effects (i.e., predictability). Moreover, despite that predictability can manifest in RT and acoustic realization, few studies examine both measures to get a better understanding of how RT and acoustics might relate to one another during phonological and phonetic encoding.

Methods. The current study attempted to explore these issues by examining the effect of high vs. low frequently-occurring monosyllabic (e.g., *Kind* 'child' vs. *Gift* 'poison') and disyllabic words (e.g., *Fehler* 'mistake' vs. *Feder* 'feather') on the acoustic vowel duration in a stressed syllable in German. Word and syllable frequency were based on CELEX and SUBTLEX-DE. Since word and syllable frequency covary in monosyllabic words, we attempted to disentangle them by using disyllabic stimuli in which syllable frequency was controlled for. Nineteen monolingual German adults (8M, 11F, mean age = 23 years) participated in a production task. In the task an object (target) was visually presented first. Then participants were instructed to click a button triggering the aural presentation of a prompt question. Participants' task was to produce an utterance incorporating the name of the object. Each target stimulus was elicited in 2 utterance positions: medial (non-final) vs. final. Prompt questions were manipulated to elicit the utterance positions of the targets. For instance, the prompt question - *Welche Feder ist rot*? 'Which feather is red?' was used to elicit the utterance-medial target in the verbal response: *Die linke Feder ist rot*. 'The left feather is red'. The prompt question – *Was malt der Künstler*? 'What does the painter draw?' was used to elicit the utterance-final target in the verbal response: *Der Künstler malt die Feder*. 'The painter draws the feather'. Depending on the prompt questions, either a single object or two versions of an identical object were visually presented. Response time (RT) was measured from the onset of the auditory prompt question to the onset of the verbal response. We expected short acoustic vowel duration and fast RT in high frequently-occurring word and syllable.

Results. Monosyllabic and disyllabic words were analyzed separately using lmer (Bates, 2015) in R (R Core, 2022), because syllable frequency was varied in the former, but controlled in the latter. Vowel duration was measured and normalized against the duration of the embedding word to derive a vowel ratio. RT was measured from the beginning of the prompt question to the beginning of the verbal response. A preliminary analysis of 5 participants showed the following patterns: (1) in monosyllabic words, vowel ratio significantly differentiated tense from lax vowels ($F = 161.89$, $df = 1$, $p < .0001$***); RT was significantly longer (slower) in response to *high frequently-occurring* than low frequently-occurring words, although the direction is counter to expectation ($F = 6.29$, $df = 1$, $p = .04$* (see **Figure 1**); (2) in disyllabic words, vowel ratio also significantly differentiated tense from lax vowels and this effect interacted with utterance position ($F = 21.4$, $df = 1$, $p < .0001$***); but no effects or interaction was observed on RT (see **Figure 2**).

Discussion. Tentatively, the different ratio patterns in monosyllabic vs. disyllabic words suggest no word-based frequency effect on disyllabic words. The word-based frequency effect on the acoustics of the monosyllabic words seems to be driven by covarying syllable-based frequency. The unexpected slower RT to frequently-occurring unit in monosyllabic words raises further questions as to the processes of retrieval or encoding in the syllabary, with implications for spoken language production planning. Interestingly, 'frequency' does not seem to influence the RT to generate the disyllabic words, although this should be taken with a grain of salt due to the possible lack of statistical power.

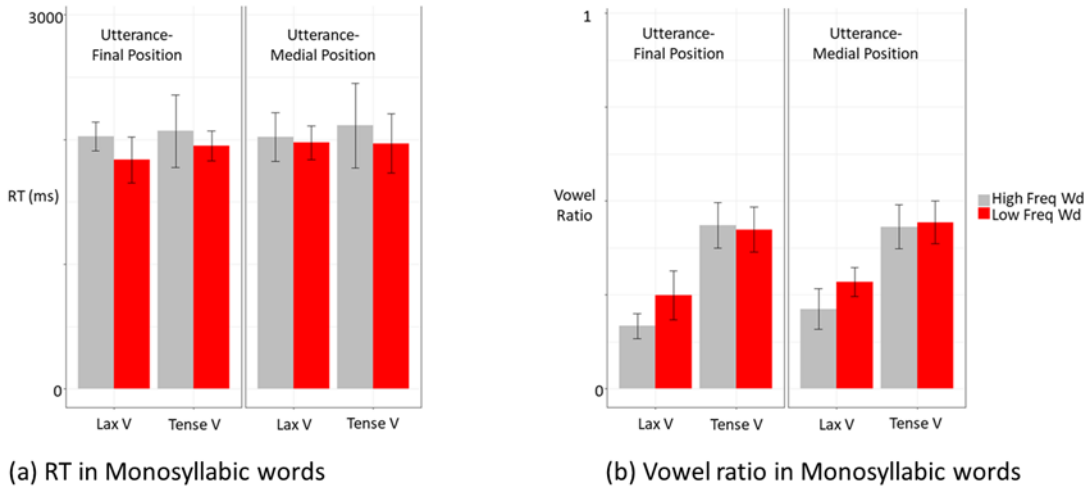


Figure 1: :(a) Mean RT and (b) mean vowel ratio to monosyllabic words containing a lax or tense vowel in final vs. medial utterance positions, with +/- 1 SD

Mean RT of disyllabic and monosyllabic words in final vs. medial utterance positions, with +/- 1 SD

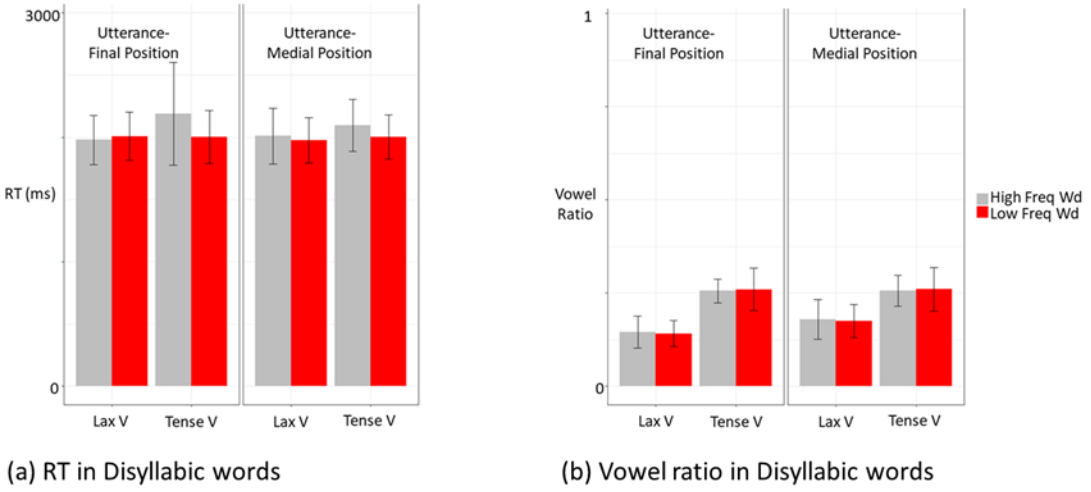


Figure 2:(a) Mean RT and (b) mean vowel ratio to disyllabic words containing a lax or tense vowel in final vs. medial utterance positions, with +/- 1 SD

References

Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, 56 (3), 349-371.

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence duration in spontaneous speech. *Language and Speech*, 47, 31-56.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Cholin, J., Levelt, W.J.M., & Schiller, N.O. (2006). Effects of syllable frequency in speech production. *Cognition*, 99, 205-235.

R Core Team (2022). R: A Language and Environment for Statistical Computing

Music in the treatment of childhood motor speech disorders: Using music to cue gestural timing

Mirjam van Tellingen[1,2], Joost Hurkmans[1], Anne Marie van der Zande[3], Hayo Terband[4], Ben Maassen[2], Roel Jonkers[2]

[1]Rehabilitation Center 'Revalidatie Friesland, Beetsterzwaag, The Netherlands,
[2]Center for Language and Cognition, University of Groningen, Groningen, The Netherlands,
[3]Rehabilitation Center 'Rijndam Revalidatie', Rotterdam, The Netherlands
[4]Department of Communication Sciences and Disorders, University of Iowa, Iowa City IA, USA

m.van.tellingen@revalidatie-friesland.nl, j.hurkmans@revalidatie-friesland.nl, avdzande@rijndam.nl, hayo-terband@uiowa.edu, b.a.m.maassen@rug.nl, r.jonkers@rug.nl

Introduction. Speech sound disorders (SSD) in children are defined as a range of difficulties in producing speech sounds and prosody due to a variety of limitations in perceptual, motor or linguistic processes (McLeod & Baker 2017). In daily life, difficulties in producing speech sounds and prosody result in reduced intelligibility, negatively affecting functional communication and participation in social situations (Hustad 2012). Diepeveen et al. (2022) showed in a recent study that more severe cases of SSD often have a combination of problems in linguistic and motor speech processes.
Treatments for motor speech disorders use principles of motor learning (Maas et al. 2014), with an articulatory-kinematic or rate/rhythm-control type approach. An example of a rate/rhythm control type approach is Speech-Music Therapy for Aphasia, a method that was originally developed for adults with aphasia and/or apraxia of speech (SMTA; De Bruijn et al. 2005). SMTA is applied in the treatment of children with motor speech sound disorders and combines speech therapy and music therapy, with the treatment being provided simultaneously by both therapists. In this treatment, target items are chosen to be both functionally relevant and fitting for the speech targets and communication goals of the individual child. The music therapist composes unique melodies that support the natural prosody of the chosen target item. During practice, musical support is phased out in a protocol that starts with singing, followed by rhythmic chanting, and speaking. During the speaking phase, the support that is given by the speech therapist is phased out, starting with simultaneous speaking, followed by imitation, and ending with response to a question (see van Tellingen et al. 2023 for a detailed description).
The use of music in the treatment of childhood motor speech disorders is supported by theories on similarities between speech and music. The first similarity concerns the overlap in neural processing of music and speech. Patel (2014) found that music training contributes to improved processing of speech through shared neural processing pathways. Expanding on this finding, Fujii and Wan (2014) hypothesise that rhythm is the working element in treatments for speech production that use musical elements. In their hypothesis, rhythm is posed as the facilitator for sound envelope processing and synchronization and entrainment to a pulse. The second similarity concerns prosody. Prosody in speech is realised through the modification of the features pitch, duration and intensity (Terband et al. 2019), which are similar to the musical parameters of melody, rhythm and dynamics (Hurkmans 2016). The third similarity is found in the timing relations of speech and music at the level of producing the elements that form phrases or melodies. In the articulatory phonology model (Browman & Goldstein 1992), timing of speech gestures in a gestural plan is expressed in relative phasing, which can be visualised in gestural scores (**Figure 1b**), showing phasing and duration for gestures in the vocal tract variables that are modulated. In a similar manner, the notes in a melody are organised in time. **Figure 1a** shows a visual representation of this phasing and duration of musical notes for multiple instruments in a musical score. The overlap in the phasing and duration of elements involved in both activities (**Figure 1b**) may contribute to the facilitatory effect of music in the rehabilitation of speech production. In the present study, we explored the effect of music on speech production by using music to provide an auditory cue for the phasing relationships of speech gestures in consonant clusters.

Methods. SMTA was evaluated in a single subject design study. The study protocol included a pretest, baseline, treatment period, post-test, follow-up period with no treatment and a follow-up test. During the 10-week treatment period, SMTA was provided twice a week in 30-minute sessions. Melodies for items including initial consonant clusters were composed including an anacrusis (pickup; a note that precedes the first beat of a measure), to serve as a cue for the realization of the cluster (see **Figure 1a** for an example of the treated item /knʉfəl/, which is Dutch for plush toy). The duration of the anacrusis was manipulated to vary the timing of sequential realisation of the consonants in the cluster and thereby decreasing or increasing difficulty. The participant in this study was a five-year-old Dutch-speaking boy with childhood apraxia of speech. Outcome measures were selected to reflect changes in intelligibility and the production of speech sounds in tasks such as picture naming, non-word imitation and spontaneous speech (see van Tellingen et al. 2023 for the

full description of this single subject study design). In the present study, we focused on the realization of initial consonant clusters in a picture naming task.

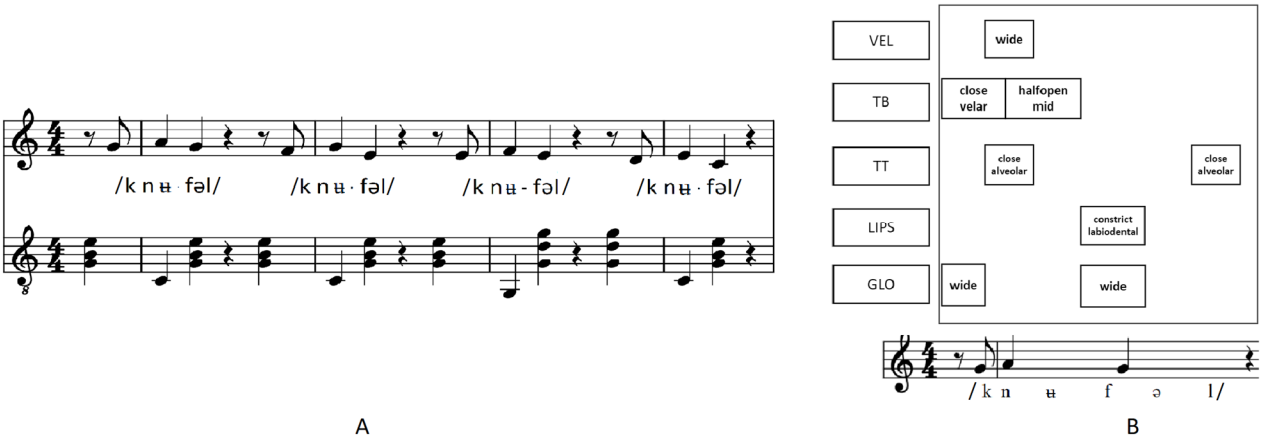


Figure 1AB: *A Musical score for SMTA melody for /knʉfəl/, B Combined scores for speech and music gestures for /knʉfəl/. VEL=velum, TB=tongue body, TT=tongue tip, GLO=glottis.*

Results. The production of initial clusters in the picture naming task improved from 62% correct at the pretest to 100% correct at the post-test (z = +5.46; with a change of z +0.5 being clinically relevant). These gains were partially maintained at follow-up, with 93% correct (z = +4.29).

Discussion. In this study similarities in the processing of phasing and duration of speech gestures and music notes were used to train the production of consonant clusters. More specifically, anacrusis was used as an auditory rhythmic cue to support the realization of gestures that need to be coordinated to produce these clusters.
In this single-subject design study, the realization of initial consonant clusters improved after treatment. These results are in line with the idea that musical structures can be used as a cue for phasing relationships of speech gestures, but replication of these results in a larger group of participants is warranted.
The effect of music in rehabilitation of motor function has been attributed to pulse entrainment (Thaut & Abiru 2010). In the present study, cues were not presented in a stable rhythm, but rather highlighted the specific phase relationship of the articulatory gestures forming consonant clusters. Therefore, the correct realization of clusters after the treatment in this study may be more in line with the concept of rhythmic tracking (Haegens 2020) than entrainment. Further research is needed to interpret clinical results in relation to the potential working mechanisms of the rhythm component in SMTA in rehabilitation of speech production.

References

Browman, C. P., & L. Goldstein (1992). "Articulatory Phonology: An Overview". In: *Phonetica*, *49*.3–4, pp. 155–180.

De Bruijn, M., T. Zielman, & J.J.S. Hurkmans (2005). "Speech-Music Therapy for Aphasia (SMTA)". Revalidatie Friesland.

Diepeveen, S., H. Terband, L. van Haaften, A.M. van de Zande, C. Megens-Huigh, B. de Swart, & B. Maassen, (2022). "Process-Oriented Profiling of Speech Sound Disorders". In: *Children*, 9.10, pp. 1502.

Fujii, S., & C.Y. Wan (2014). "The role of rhythm in speech and language rehabilitation: The SEP hypothesis". In: *Frontiers in Human Neuroscience*, 8, Article 777.

Haegens, S. (2020). "Entrainment revisited: A commentary on Meyer, Sun, and Martin (2020)". In: *Language, Cognition and Neuroscience*, 35.9, pp. 1119–1123.

Hurkmans, J. J. S. (2016). "The treatment of apraxia of speech". Rijksuniversiteit Groningen.

Hustad, K. (2012). "Speech Intelligibility in Children With Speech Disorders". In: *Perspectives on Language Learning and Education*, 19.1, Article 1.

Maas, E., C.E. Gildersleeve-Neumann, K.J. Jakielski, & R. Stoeckel (2014). "Motor-Based Intervention Protocols in Treatment of Childhood Apraxia of Speech (CAS)". In: *Current Developmental Disorders Reports*, 1.3, Article 3.

McLeod, S., & E. Baker (2017). "Children's Speech: An Evidence-Based Approach to Assessment and Intervention". Pearson.

Patel, A. D. (2014). "Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis". In: *Hearing Research*, 308, pp. 98–108.

Terband, H., A Namasivayam, E. Maas, F. van Brenk, M.L. Mailend, S. Diepeveen, P. van Lieshout & B. Maassen. (2019). "Assessment of Childhood Apraxia of Speech: A Review/Tutorial of Objective Measurement Techniques". In: *Journal of Speech, Language, and Hearing Research*, 62.8S, Article 8S.

Thaut, M. H., & M. Abiru (2010). "Rhythmic Auditory Stimulation in Rehabilitation of Movement Disorders: A Review Of Current Research". In: *Music Perception*, 27.4, pp. 263–269.

van Tellingen, M., J. Hurkmans, H. Terband, A.M. van de Zande, B. Maassen, & R. Jonkers (2023). "Speech and Music Therapy in the Treatment of Childhood Apraxia of Speech: An Introduction and a Case Study". In: *Journal of Speech, Language, and Hearing Research*, pp. 1–19.

Merging verb forms with "ich" to *enchaînement consonantique* in German

Jürgen Trouvain[1], Christine Mooshammer[2], Malte Belz[2], Robert Lange[2]

[1]Language Science and Technology, Saarland University, Saarbrücken, Germany
[2]German Studies and Linguistics, Humboldt-Universität zu Berlin, Germany
trouvain@lst.uni-saarland.de,
christine.mooshammer|malte.belz|robert.lange@hu-berlin.de

Introduction. In German, the personal pronoun of the first person, singular, "ich" ("I") is the second most frequent word in conversations (Brackhane 2022). Being a function word and a word with an ultra-high frequency of occurrence leads to a very high predictability (Jurafsky et al. 2001). Consequently, it can be assumed that "ich" is not located in an accented position. Words in unaccented locations are usually produced with a comparably shorter duration and a higher degree of phonetic reduction (Kohler 1990). However, a further but so far neglected effect is resyllabification. When the preceding word ends with a consonant, and the vowel in "ich" is not deleted, it is very likely that the word sequence "weil ich" ("because I") would be syllabified as "wei.lich" or "wenn ich" to "wen.nich" ("if I" or "when I" – this example is with an ambisyllabic consonant due to the preceding short/lax vowel). This cross-word sandhi effect is known as *enchaînement consonantique* in French (e.g. Oh et al. 2023) but has not yet been described for German to the best of our knowledge.
An optimal test bed for such an investigation are finite verb forms located before the personal pronoun "ich", a word sequence often required in German syntax. Full forms of verbs ending on a consonant are restricted to few verbs such as "kann" ("can") and "schrieb" ("wrote"), most verbs have a schwa in their orthographic form, as in "habe ich". If the schwa is realised, then the sandhi would not occur. However, Kohler (2001) finds that schwa is not realised in more than 68% of cases when followed by a vowel. Similarly, Wesener (1999) reports 76% schwa deletions for verbs. After schwa deletion, the output will be *enchaînement*. Compare e.g. the full form "nehme ich" ("take I") => "neh.me.ich" with the schwa-deleted form "nehm ich" => "neh.mich".
Schwa deletion potentially leads to final devoicing if the then-final consonant is a voiced obstruent: final devoicing (*Auslautverhärtung*) takes place when the voiced obstruent remains in coda position as in the imperative forms "glaub mir" ("believe me") or "sag mal" ("tell me"). However, schwa-deleted verb forms followed by "ich" would undergo *enchaînement* but with an unclear treatment of *Auslautverhärtung* when there is a lenis obstruent at the syllable boundary. Would "sag ich" ("say I") result in [za:kɪç] or in [za:gɪç], i.e. with final devoicing or without? Final devoicing would suggest that no resyllabification has happened. Verbs that have a tense/long vowel before the schwa as in "gehe ich" ("go I") would not undergo a resyllabification after schwa-deletion. Consequently, we can expect different outputs for verbs before "ich": either with or without schwa, and for the forms without schwa: either with resyllabification (and no glottal stop at the syllable boundary) or without resyllabification (and glottal stop). Compare some possible examples with "habe ich" in broad phonetic transcription: (1) [ha:**bəʔ**ɪç] vs. (2) [ha:**bə**ɪç] vs. (3) [ha:**pʔ**ɪç] vs (4) [ha:**b**ɪç].
The aim of this study is to explore verbs followed by "ich" in German spontaneous speech of the type "habe ich" (consonant before possibly deleted schwa) or "kann ich" (without underlying schwa). How often can we observe schwa deletion in verb forms before "ich"? How often do verb forms without schwa (both types) before "ich" lead to a subsequent *enchaînement*?

Methods. The inspected data were taken from the Corpus of Non-Native Addressee Register (CoNNAR) (Lüdeling et al 2023; Terada et al. 2023). The corpus was designed to investigate the intra-individual linguistic variation when addressing different interlocutors in free and task-based conversations, here German native speakers or learners of German as a foreign language. The analysed data contain 120 conversations between 20 participants and eight instructed interlocutors whereby only the data of the participants are considered (85,949 cleaned word tokens, i.e. 6.8 hours of articulation time). Data annotation was performed on different tiers, e.g. on the word and the phone level, with all labels manually checked by auditory impression plus visual inspections of the speech signal. For this study, all sequences of finite verbs followed by "ich" were extracted. Usually, a sequence of verb+ich leads to a verb phrase. In few cases, such a sequence could also contain a syntactic boundary. Regarding the segmental realisations we only looked at schwa, glottal stops and resyllabifications at the (potential) sandhi location.

Results. The personal pronoun "ich" occurs 2,200 times in the inspected sub-corpus. In 37.1% of these cases it is preceded by a finite verb. The vast majority of type "habe ich" do not show schwa realisation making them candidates for resyllabification, if a consonant is present (see Figure 1). Virtually **all** of schwa-less verb forms (of both types) do apply resyllabification, i.e. without a glottal stop and without devoicing of then syllable-final consonants, corresponding to example 4 from the type [ha:**b**ɪç].

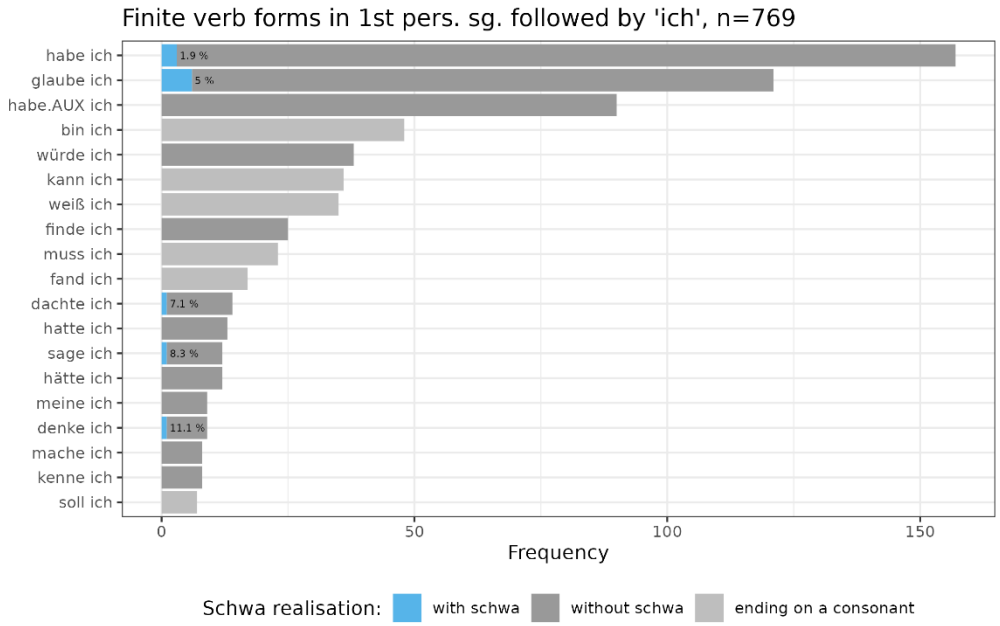


Figure 1: Frequency of inspected finite verb forms plus "ich" with more than 5 occurrences in CoNNAR. Full forms with schwa in blue (n=14), schwa-deleted forms in gray (n=579), no schwa forms in light gray (n=176).

Discussion. As expected, schwa deletion is common practice in realisations of finite verb forms in our data of German spontaneous speech. The option for a resyllabification of the type of e*nchaînement* seems to represent the default case. Since this phonological phenomenon has been neglected so far in the phonology of German, the presented finding provokes questions regarding its morpho-phonological analysis and explanation. Schiering (2002) gives an overview of different accounts in German phonology of cliticisation. However, there is no general agreement on how to regard verb forms with deleted schwa and following pronouns with an unreduced vowel such as "ich", some see it as a clitic group, similar to "hat es" => "hat's" ("there's"), others see it as an affix-like structure.

Conclusion. *Enchaînement consonantique* exists in French but also in German, but it was neglected so far as a phonological phenomenon. Interestingly, it is a highly frequent phenomenon, that however occurs in prosodically unaccented locations, so it might be not very prominent in the perception of linguists. Although German belongs to the languages that are better described for connected speech processes than others, there are still phenomena to be described in more detail, and not all of them are linked to fast articulation as the sometimes-used term "allegro rules" imply. Annotated corpora of various speech styles and registers are very helpful in unveiling those spots. Examples of applying this type of research concern language teaching. Since e*nchaînement consonantique* also happens in their native language, German learners of French could be made more sensitive to resyllabification effects in French, including *liaison*. Likewise, for learners of German it would be an important illustration of how unaccented passages of fluent speech, i.e. the majority of words, are produced in conversations.

Parts of this research were funded by the Deutsche Forschungsgemeinschaft – SFB 1412, 416591334.

References

Brackhane, F. (2022). Beobachtungen zu Frequenz und Funktionen von *ja* in deutscher Spontansprache. *Deutsche Sprache*, 4/2022, pp. 335–363.

Jurafsky, D., Bell, A., Gregory, M. & Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production, in J. Bybee & P. Hopper (Eds.), Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins, 229–254.

Kohler, K. J. (1990). Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In W. Hardcastle & A. Marchal (Eds.), Speech production and speech modelling. Dordrecht: Kluwer Academic. 69–92.

Kohler, K. J., Rodgers, J. E. J. (2001). Schwa deletion in German read and spontaneous speech. In: Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK) 35. pp. 97–123.

Lüdeling, A. Mooshammer, Ch., Lange, R., Sell, B. M., & Terada, M. (2023). Corpus of Non-Native Addressee Register (CoNNAR). Version 1. https://rs.cms.hu-berlin.de/phon/pages/home.php

Oh, S, Fougeron, C., Buech, Ph., Hermes, A. (2023). CV coordination: the case of enchaînement and liaison in French. ICPhS Prague, pp. 1137–1141.

Schiering, R. (2002). Klitisierung von Pronomina und Artikelformen. Eine empirische Untersuchung am Beispiel des Ruhrdeutschen. Arbeitspapier 44, Institut für Sprachwissenschaft, Universität zu Köln.

Terada, M., Sell, B. M., Lange, R., Müller, M., & Belz, M. (2023). Documentation and annotation guidelines of CoNNAR Version 1. *Register Aspects of Language in Situation (REALIS)*. 2(6), pp. 1–32, doi=https://doi.org/10.18452/27898.

Wesener, Th. (1999). The phonetics of function words in German spontaneous speech. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK) 34, pp. 323–373.

Use of Natural Anchors for Improving Rater Reliability in Dysarthria Assessment: An Exploratory Study

Mili Kuruvilla-Dugdale[1], Thushani Umesha Munasinghe[1], Deepthi Crasta[1], Kaila Stipancic[2], Amy Anil[1]

[1]University of Iowa, Iowa, USA

[2]University at Buffalo, USA

mili-kuruvilla-dugdale@uiowa.edu, klstip@buffalo.edu, thushani-munasinghe@uiowa.edu, deepthi-crasta@uiowa.edu, amy-anil@uiowa.edu

Introduction. When listeners rate speech samples produced by talkers with dysarthria, they often rely on idiosyncratic internal standards and compare the presented sample to an internal reference (Kreiman et al., 1993). Internal standards vary among listeners based on their experience with dysarthria, and differential attention to dysarthric features. These standards are also influenced by memory and context, causing them to become unstable and to drift over time (Gerratt et al., 1993). As a result, low rater reliability is a central issue when performing auditory-perceptual judgments for subtyping dysarthria and quantifying severity (Stipancic et al., 2023). One way to minimize the high variability among raters is to substitute listeners' internal standards with stable external anchors (Awan & Lawson, 2009; Chan & Yiu, 2002).

External anchors have previously been implemented with auditory-perceptual scaling methods (e.g., Awan & Lawson, 2009). Among the available rating scales, two are preferred for assessment: interval scales, which use a predetermined set of categories or numbers to assign to stimuli (e.g., equal appearing interval scale [EAI]), and ratio scales, which require the assignment of numerical values proportional to the perceived ratio of the reference stimulus (e.g., direct magnitude estimation [DME]). When using external anchors with EAI, the experimenter can assign an anchor to each scale point to assist raters with their judgments. Similarly, DME employs a modulus (i.e., an anchor), where listeners rate stimuli in comparison to the standard anchor of moderate severity selected by the experimenter.

Several voice studies, but no dysarthria studies, have explored the use of anchors by manipulating various components of anchor implementation, such as anchor type (synthetic and natural) (Santos et al., 2021) and modality (auditory, visual, or both) (Awan & Lawson, 2009). Reported findings suggest improved inter-rater reliability with anchor use, particularly when anchors are paired with auditory training (Wong et al., 2021). Although previous research has employed DME for global ratings of intelligibility and severity in speakers with dysarthria (Southwood, 1996; Weismer et al., 2002), none of these studies have investigated the effects of anchors on DME reliability. To this end, the present study aimed to compare rater reliability with and without anchors when using EAI and DME to scale features of hypokinetic dysarthria including overall speech impairment severity, articulatory imprecision, reduced loudness, short rushes of speech, and monotony.

Methods. Sixty-eight speech samples recorded from people with Parkinson's disease (PD; $n = 43$) and healthy controls ($n = 25$) were used for the ratings; 14 samples were re-rated to allow for calculation of intrarater reliability. Listener recruitment is ongoing, but so far, 14 non-expert healthy participants (*mean age* = 26.5 years, *SD* = 3.55) have rated the speech samples using EAI and DME scales. We anticipate 40 listeners will have completed the study at the time of the conference in May 2024. The Speech Intelligibility Test (SIT; Yorkston et al., 2007) sentence selected from each PD speaker included several of the target hypokinetic speech features. The perceptual ratings were performed over four sessions (i.e., EAI with anchors; EAI without anchors; DME with anchors; DME without anchors) that were each about one week apart and lasted approximately an hour each. Scale, anchor condition, and sample order were randomized across participants.

For the two EAI sessions, listeners were provided with a 5-point EAI scale either with or without anchors and were asked to rate the five features after listening to each sample twice. Feature definitions were provided to each listener. The 5-point EAI scale had the following intervals: 1=typical, 2=mild, 3=moderate, 4=severe, and 5=profound.

For the two DME sessions, the listeners were asked to rate the speech features with or without an anchor. For the former task, the experimenter selected an anchor for each feature, and listeners were asked to rate all samples in comparison to the given anchor. The anchor represented moderate severity for each feature and was given a score of 100. For the latter task, listeners were asked to assign any number to their first stimulus; all subsequent samples were rated relative to this reference and given a comparative score. For example, if the sample was half as severe as the anchor, a score of 50 was recommended.

Interrater reliability was estimated through intraclass correlation coefficients (ICCs) for each scaling method (i.e., EAI and DME) and anchor condition (i.e., with and without anchors). ICCs and their 95% confidence intervals (CIs) were calculated using SPSS statistical package version 28 (SPSS Inc., Chicago, IL) based on single- and average-measures consistency, 2-way mixed-effects model with 14 raters across 68 samples. A single measure ICC is based on a single measurement, and the average measure ICC is based on the average measurements of more than one observer. Intrarater reliability was judged using Spearman's correlation coefficients. Only the EAI results are included here; data analysis for DME is in progress and will be included in the conference presentation.

Results. The inter and intrarater reliability results for EAI across the five features are summarized in Table 1. Overall, for interrater reliability, there was an increase in both single and average measures ICC for all features when anchors were used, compared to when anchors were not used. The average measures ICC is highly acceptable (good or excellent) for both anchor conditions. In contrast, the single measures ICC for all features ranged between poor and moderate. A

meaningful change in reliability (i.e., switch to a higher reliability category) from good to excellent was noted for the average measures ICC of short rushes of speech when anchors were used. Similarly, single measures ICC of overall severity and reduced loudness changed from poor to moderate with the use of anchors.

Intrarater reliability also improved with anchor use for all features except for monotony, where a slight decrease in reliability was noted when scaling with anchors than without anchors. A meaningful improvement in reliability was noted for articulatory imprecision and reduced loudness, but not for overall severity or short rushes of speech despite increases in reliability statistics with the use of anchors for these features.

Table 1. Rater reliability for equal appearing interval (EAI) ratings of all features with and without anchors.

Speech Feature	Interrater Reliability Intraclass Correlation Coefficient (CIs)			Intrarater Reliability Spearman's Correlation Coefficient (CIs)	
	Measure	No Anchors	With Anchors	No Anchors	With Anchors
Overall severity	Single	0.492 (.403-.593)	0.587 (.501-.680)	0.760 (.692-.815)	0.802 (.743-.848)
	Average	0.931 (.904-.953)	0.952 (.934-.967)		
Articulatory imprecision	Single	0.513 (.425-.613)	0.597 (.512-.689)	0.642 (.548-.719)	0.766 (.699-.820)
	Average	0.937 (.912-.957)	0.954 (.936-.969)		
Reduced loudness	Single	0.423 (.337-.526)	0.520 (.432-.619)	0.680 (.594-.751)	0.790 (.728-.839)
	Average	0.911 (.877-.940)	0.938 (.914-.958)		
Short rushes of speech	Single	0.295 (.219-.393)	0.410 (.324-.513)	0.550 (.441-.643)	0.651 (.559-.727)
	Average	0.854 (.797-.901)	0.907 (.870-.936)		
Monotony	Single	0.433 (.346-.536)	0.495 (.407-.596)	0.666 (.577-.739)	0.654 (.562-.729)
	Average	0.914 (.881-.942)	0.932 (.906-.954)		

Note. ICC values less than 0.5 are indicative of poor reliability; values between 0.5 and 0.75 indicate moderate reliability; values between 0.75 and 0.9 indicate good reliability; and values greater than 0.90 indicate excellent reliability. Spearman's correlation coefficients from 0.99-0.70 suggest a strong association; values between 0.69-0.50 indicate an average association; and values from 0.49-0.01 indicate a weak association.

Discussion. Our preliminary findings suggest that there are benefits to using anchors with interval and ratio scales during dysarthria assessment. Voice studies have reported similar improvements in reliability when external anchors are combined with rater training. However, these studies also show increased intra- and inter-rater variability when anchors were used without training, suggesting limited use for anchors alone (Chan & Yiu, 2006). The current findings suggest that the benefits to reliability from anchor use vary based on the feature being rated. For intrarater reliability, a meaningful change was observed only for articulatory imprecision and reduced loudness, whereas for interrater reliability, a meaningful change was observed for overall severity (single measures ICC), reduced loudness (single measures ICC), and short rushes of speech (average measures ICC). Despite this improvement in interrater reliability, the moderate reliability observed for overall severity and reduced loudness when using anchors is still insufficient for clinical purposes, which contrasts with the average measures ICC for both anchor conditions, which are highly acceptable. Hayen and colleagues (2007) emphasize that the average should not be used when determining ICCs unless there are specific situations where averaged ratings apply. In clinical dysarthria assessments, a single listener typically rates a single speaker, which would necessitate interpreting reliability using a single measure ICC. Overall, researchers and clinicians should consider using external anchors for scaling features of dysarthric speech even if they are highly experienced as internal standards can shift over time. While the current results are limited to EAI scales, the DME results will be presented at the conference to emphasize which scales and features pair well with anchors to improve reliability, and the implications of the findings.

References

Awan, S. N., & Lawson, L. L. (2009). The Effect of Anchor Modality on the Reliability of Vocal Severity Ratings. *Journal of Voice*, 23(3), 341–352.

Chan, K. M. K., & Yiu, E. M.-L. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research*, 45(1), 111–126.

Eadie, T. L., & Kapsner-Smith, M. (2011). The Effect of Listener Experience and Anchors on Judgments of Dysphonia. *Journal of Speech, Language & Hearing Research*, 54(2), 430–447.

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. S. (1993). Comparing Internal and External Standards in Voice Quality Judgments. *Journal of Speech, Language, and Hearing Research*, 36(1), 14–20.

Hayen, A., Dennis, R. J., & Finch, C. F. (2007). Determining the intra- and inter-observer reliability of screening tools used in sports injury research. *Journal of Science and Medicine in Sport*, *10*(4), 201–210.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research*, 36(1), 21–40.

Santos, P. C. M. D., Vieira, M. N., Sansão, J. P. H., & Gama, A. C. C. (2021). Effect of synthesized voice anchors on auditory-perceptual voice evaluation. *CoDAS*, 33(1), e20190197.

Southwood, M. H. (1996). Direct magnitude estimation and interval scaling of naturalness and bizarreness of the dysarthria associated with amyotrophic lateral sclerosis. *Journal of Medical Speech-Language Pathology*, 4(1), 13–25.

Stipancic, K. L., Golzy, M., Zhao, Y., Pinkerton, L., Rohl, A., & Kuruvilla-Dugdale, M. (2023). Improving Perceptual Speech Ratings: The Effects of Auditory Training on Judgments of Dysarthric Speech. *Journal of Speech, Language, and Hearing Research*, 66(11), 4236–4258.

Yorkston, K. M., Beukelman, D., & Hakel, M. (2007). *Speech Intelligibility Test for Windows [Computer software]*. Madonna Rehabilitation Hospital.

Weismer, G., & Laures, J. S. (2002). Direct Magnitude Estimates of Speech Intelligibility in Dysarthria: Effects of a Chosen Standard. *Journal of Speech, Language, and Hearing Research,* 45(3), 421–433.

Muscle synergies in the production of stop consonants with increasing intensity

Maëva Garnier, Marion Léger, Julien Frère

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
maeva.garnier@gipsa-lab.fr

Introduction. Speech production involves a multitude of muscles, including about ten around the lips (Hjortsjö, 1970). Various studies have already documented their anatomy and involvement in different orofacial movements (Folkins, 1976; Lapatki et al., 2006). It has been established, in particular, that very few of these muscles are recruited in isolation, but that most of them rather function in synergy (Ekman et Friesen, 1976). Several studies have characterized the orofacial muscle patterns involved in the production of different speech sounds (Schumann et al., 2010; Eskes et al., 2017 ; Wand et al., 2013), at a "comfortable" level corresponding to a usual situation of face-to-face interaction. The aim of this study is to explore how these orofacial muscle activation patterns vary with articulatory effort, for speech produced from a whispered (hypo-articulated) to a shouted (hyper-articulated) level. In particular, the question arises as to whether muscle synergies are relatively conserved, or significantly reorganized, with increasing articulation effort (by analogy, when switching from walking to running; Cappellini, et al., 2006).

Methods. This study is based on the FullStop database (Cattelain 2019), in which 20 French speakers were recorded while producing repetitions of non-words /laCV/, with C={p, b} and V={a, i} in modal phonation, at comfortable and fast rates, and with 5 increasing levels of intensity (defined subjectively by the speaker's sense of effort : from murmur to shout). The database contains the audio signal of these productions (calibrated in dB SPL), synchronized with other physiological signals, including variations in lip aperture, extracted from high-speed video images (200 f/s) and the electromyographic signal (EMG) of 5 lip and neck muscles: the Orbicularis Oris Inferior (OOI) and Superior (OOS), the Depressor Labii Inferior (DLI), as well as the Mentalis (MNT) and Digastric (DIG) muscles.
Non-negative matrix factorization analyses (NNMF) (Seung and Lee, 2001) were performed on all EMG signals, normalized both temporally (to the duration of each word) and in amplitude (to the maximum of the RMS envelope of each word), in order to identify groups of muscles that work in synergy, and quantify the relative contribution of the 5 investigated muscles to each of these synergies.
To interpret these synergies in terms of motor control, the lip movement involved in the production of the labial stops was decomposed into four phases, defined and detected from the variations in lip aperture: P1-lip closing; P2-lip compression during occlusion; P3-lip decompression during occlusion and P4-lip re-opening after the occlusion release.
Six descriptors were also measured from each extracted synergy : 1- the time-to-peak of muscle synergy activation (in % of the total word duration), 2- the full width at half maximum value (FWHM, in % of the total word duration) that quantifies duration of muscle synergies activation (As EMG signals were normalized in amplitude prior to synergies extraction, activity levels were not taken into account here) and 3- its average activity level within each phase (P1, P2, P3 or P4) (To this end, we weighted the non-normalized RMS envelopes with the vector values of each muscle synergy (cross product)).
First, the percentage of productions for which that time-to-peak value occurred within the phase P1, P2, P3 or P4 was computed. Generalized linear mixed models were then used to explore the influence of vowel context (a vs. i), consonant voicing (p vs. b) or speech rate (normal vs. fast) on time-to-peak values, FWHM and level of muscle synergy activity.
Finally, we ran correlation analysis to explore the extent to which the activity level of these muscle synergies in the different phases of movement predicted certain kinematic characteristics (degree of lip compression during the occlusion, lip opening velocity at occlusion release), as well as the acoustic intensity of the resulting consonantal noise.

Results. The NNMF analyses enabled us to identify 2 muscle synergies underlying the production of labial stops (cf. Figure 1): The first synergy mainly involves the OOS and MNT, and shows maximum activation during the first half of the movement, corresponding more precisely to the P1 and P2 phases of lip closure and compression. The second synergy mainly involves the DLI, as well as the other muscles of the lower face (OOI, DIG) and the MNT, and shows maximum activation in the middle of the movement, corresponding more precisely to the P3 phase of lip decompression just before the occlusion release.
The same two muscle synergies were found for the production of consonants /p/ as /b/, in context /i/ as /a/, for both comfortable and fast speech rates. However, speech rate had a significant effect on the characteristics of these two synergies, with an earlier activation (6.97 and 3.56% of word duration earlier for synergies 1 and 2, respectively), a longer activation (+3.1 and +9.2% of word duration for synergies 1 and 2, respectively) and a higher level of activity for both synergies over the P1 to P4 phases in fast speech than in comfortable speech (cf. Figure 2). In contrast, vowel context and consonant voicing showed no significant effect on either muscle synergy.

Finally, a significant correlation was observed, for the syllables /pa/ produced at comfortable rate, between the degree of lip compression and the activation level of the 1st synergy in phases P1 et P2 (R=0.43 et R=0.45, respectively), between the lip reopening velocity and the activation level of the 2nd synergy in phase P3 (R=0.72), and between the burst intensity and the activation of the 1st synergy in phases P1 et P2 (R=0.27 et R=0.30, respectively), and that of the 2nd synergy in phase P3 (R=0.27).

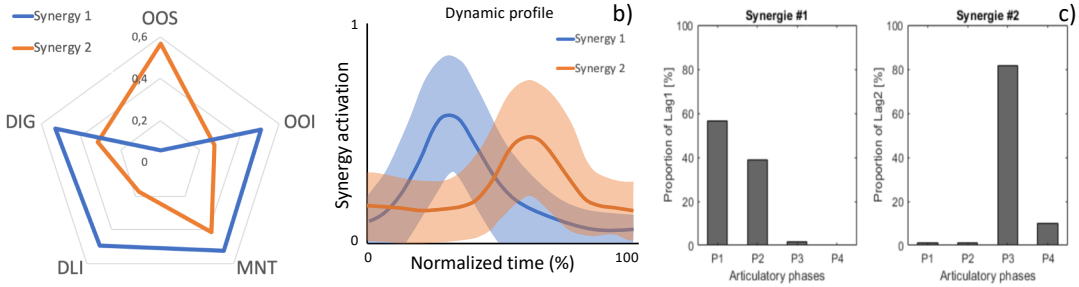


Figure 1. *Representation of the two muscular synergies observed during the production of labial stops, with for each (a) the relative weight of activity of the 5 muscles considered, (b) the temporal pattern of their activation over the word interval, and (c) the number of productions for which the maximum activation of the synergies was observed during the P1, P2, P3 or P4 phases of the movement.*

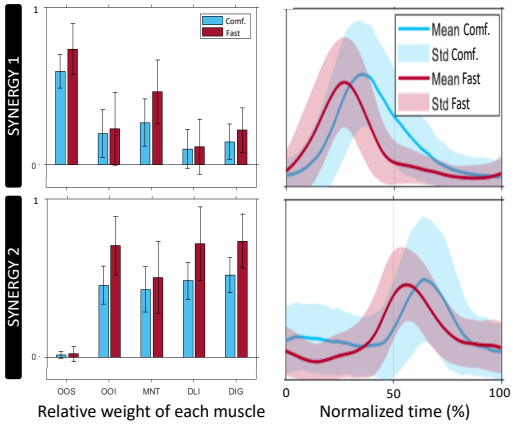


Figure 2. *Influence of speech rate on the relative contribution of different muscles in the two muscle synergies, and their time activation pattern.*

Discussion. The co-activation of the OOS and MNT during the lip compression movement is coherent and may well reflect a synergetic functioning of these two muscles. For the second synergy, activation of the DLI during the lip reopening phase is very coherent, while that of the MNT and the OOI is more unexpected, and could in fact correspond to the activity of the DLI, due to a crosstalk phenomenon. The production of labial stops did not show significant reorganization in muscle coordination, depending on the vowel context or consonant voicing. Only speech rate affected timing and level of activity of the muscle synergies, but composition of the muscle groups remained consistent. These results are consistent with the robustness of the central command found in other movements, for which the main degree of modulation is found at the temporal level (Cappelini et al. 2006). Finally, lip muscle activity predicts relatively well the movement kinematics (degree of interlip compression, and lip re-opening velocity), whereas it is less predictive of variations in burst intensity. This supports the idea that variations in burst intensity are not only controlled by lip articulation, but also by other articulatory and aerodynamic variations.

References

Cappellini, G., Ivanenko, Y. P., Poppele, R. E., & Lacquaniti, F. (2006). Motor patterns in human walking and running. *Journal of Neurophysiology*, 95(6), 3426-3437.

Cattelain, T. (2019) "Production des consonnes plosives du Français : du contrôle des bruit de plosion", Ph.D thesis, Université Grenoble Alpes.

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1), 56-75

Eskes, M., van Alphen, M. J., Balm, A. J., Smeele, L. E., Brandsma, D., & van der Heijden, F. (2017). Predicting 3D lip shapes using facial surface EMG. *PLoS One,* 12(4), e0175025.

Folkins, J. W. (1978). Lower lip displacement during in-vivo stimulation of human labial muscles. *Archives of oral biology*, 23(3), 195-202.

Hjortsjö, C. H. (1970). Man's Face and Mimic. *Language*.

Lapatki, B. G., Oostenveld, R., Van Dijk, J. P., Jonas, I. E., Zwarts, M. J., & Stegeman, D. F. (2006). Topographical characteristics of motor units of the lower facial musculature revealed by means of high-density surface EMG. *Journal of neurophysiology*, 95(1), 342-354.

Schumann, N. P., Bongers, K., Guntinas-Lichius, O., & Scholle, H. C. (2010). Facial muscle activation patterns in healthy male humans: A multi-channel surface EMG study. *Journal of neuroscience methods*, 187(1), 120-128.

Wand, M., Schulte, C., Janke, M., & Schultz, T. (2013, February). Array-based Electromyographic Silent Speech Interface. In Biosignals (pp. 89-96).

Effects of Aging on /s/ in Spontaneous Speech Using Neural Networks and Phonetic Measures

Orane Dufour, Anne Hermes, Cédric Gendrot

Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne Nouvelle, Paris, France
orane.dufour@sorbonne-nouvelle.fr, anne.hermes@sorbonne-nouvelle.fr,
cedric.gendrot@sorbonne-nouvelle.fr

Introduction. As the global population continues to age, understanding the effects of aging on various aspects of human communication becomes increasingly crucial. However, despite its significance, the impact of aging on spontaneous speech remains a relatively understudied area, with limited research comprehensively (i.e. on phonemically controlled specific read corpora) exploring the specific changes that occur in the aging population. This research proposal aims to investigate the effects of aging on uncontrolled spontaneous speech through a more substantial volume of data. This study examines spontaneous speech data of about 35 hours of speech from different age groups, ranging from 20 to 89 years of age. For that purpose, we will use a combination of phonetic measurements and features from neural networks to model aging effects on /s/. Aging can lead to changes in the spectral frequency and duration of the /s/ sound in that aging entails a slowing down in speech (Amerman & Parnell 1992, Hermes et al. 2018, Taylor et al. 2020), plus a decline in COG (stronger for males than females; Taylor et al. 2020), and sex-dependent aging effects on skewness. As Taylor et al. (2020:647) stated "older speakers may produce fricative spectra that differ from those of younger adults."

Methods. The data are taken from the ESLO[1] corpus and are separated in two different corpora:
(a) a transversal corpus: recordings of semi-prepared interviews conducted by researchers on 75 speakers from the module "entretiens" (from 21 to 89 years) and nine speakers from "entretiens jeunes" (from 20 to 25 years). See table 1 for more details.
(b) a longitudinal corpus: The module "diachronie" contains recordings of seven speakers which were recorded twice with 40 years apart (around 20-29 for the first recording) with a total recording time of 10 hours equally split between for both recordings.
The data have been orthographically transcribed at the phrase level.[2] As shown in Figure 1, sequences were then automatically aligned with the Montreal Forced Aligner (MFA)[3]. The alignment of speech data with their corresponding transcriptions has been compared to a manual alignment of small excerpts with a kappa score of 0.8 at 10 ms accuracy. For this preliminary study, we chose to focus on /s/ for its rather long duration and accuracy in the phonemic segmentation. The acoustic analysis was composed of four spectral moments (center of gravity, standard deviation, skewness, kurtosis) and duration. Measurements were taken in the median part of the consonant.

Results. /s/ were extracted randomly from the corpus and a minimum of 727 occurrences per speaker was chosen for the training. A preliminary analysis showed results in line with the literature (i.e., longer durations, lower COG, higher skewness for older speakers), with however an important inter-speaker variation, revealing a global aging effect, nonetheless, several speakers have opposite results. We decided to set out on a different strategy and to start from a perceptual point of view with the first author objectively distinguishing between two categories: younger vs. older. The two extremes (8 male speakers, 8 female speakers, 19% of the dataset) were chosen for extracting spectrograms of /s/, that were firstly used in a CNN and also analyzed acoustically. By separating female and male speakers, and younger vs. older, CNN classification reached 88% in female speakers and 72% for male speakers. When inspecting CNN classifications, we show that /s/ of short duration (< 70ms) have 10% to 15% less classification accuracy, and that /s/ with higher COG, and lower skewness show significantly better classification results. The /s/ with strong VCV coarticulation revealed voiced information in the lower frequencies and lowered the cog values. These occurrences had significantly 20% less correct classifications.
The /s/ used in this acoustic analysis correspond to the dataset used for the CNN test. We are aware that this information is far less representative of the production than the full spectrograms used in the CNN but we aim at showing the acoustic particularities of the /s/ that were poorly or successfully classified appear to show a decline in COG, aligning with findings reported in existing literature. The acoustic analysis revealed a marked distinction between younger and older speakers in terms of center of gravity (COG) metrics. Figure 2 shows the results for the duration and the COG according to sex and age. With advancing age, thas 'younger' exceeded those deemed 'older' for both groups. In contrast,

[1] http://eslo.huma-num.fr/
[2] http://eslo.huma-num.fr/images/eslo/pdf/GUIDE_TRANSCRIPTEUR_V2_Fevrier2010.pdf
[3] https://montreal-forced-aligner.readthedocs.io/en/latest/#

analyses of skewness, kurtosis, and standard deviation did not reveal any discernible patterns. As for the duration metric, female's /s/ predicted as 'older' did not consistently exhibit greater length compared to those predicted as 'younger'. In the case of male speakers, the duration of /s/ categorized as 'older' showed a significant increase, corroborating trends documented in prior studies reporting a slowing down in speech with age (see Figure 2).

Discussion. Attempts were made at taking into account three age groups (younger, mid and older) and the whole dataset, but classifications fell considerably with the 'mid' group failing to receive correct classifications. We are still performing new classifications and using different spectrogram and acoustic representations that will be presented at the ISSP conference. In the next future, these results will be compared with other phonemes, and will be extended to syllables

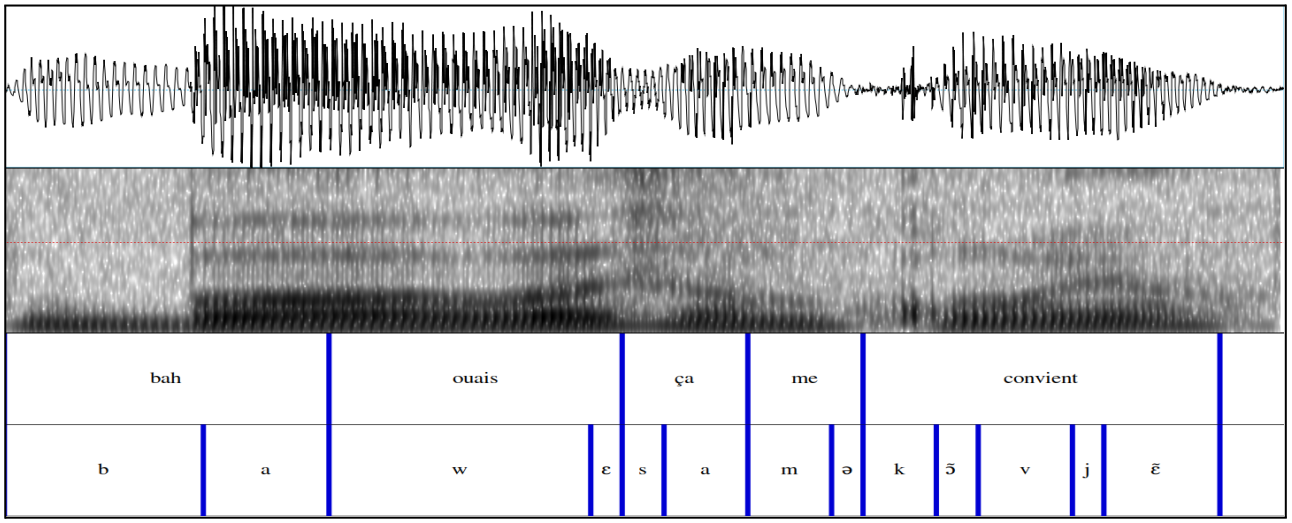
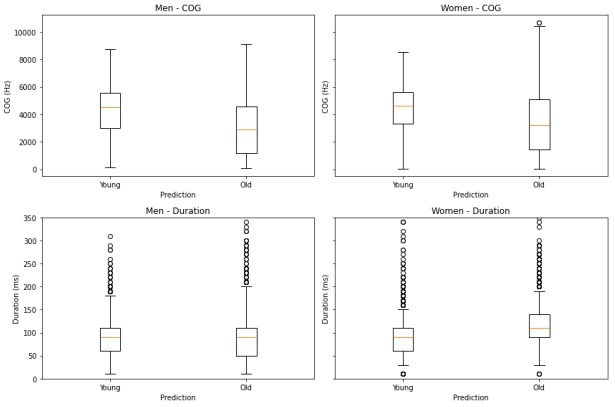


Figure 1: *MFA alignment from speaker ESLO2_ENT_1001, 1.44s*

Table 1: *Speakers age, number of speakers and duration of the recordings in the ESLO transversal corpus.*

Age	number of speakers	duration (hh:mm:ss)
20-29	20	06:36:42
30-39	16	06:10:01
40-49	14	06:19:17
50-59	12	05:29:48
60-69	13	06:07:16
70-79	3	01:15:44
80-89	6	02:37:38

Figure 2: *Duration and cog measurements of /s/ according to sex and age categories.*



References

Amerman, J. D., & Parnell, M. M. (1992). Speech timing strategies in elderly adults. *Journal of Phonetics*, 20(1), 65-76.

Hermes, A., Mertens, J., & Mücke, D. (2018). Age-related Effects on Sensorimotor Control of Speech Production. In *INTERSPEECH* (pp. 1526-1530).

Taylor, S., Dromey, C., Nissen, S. L., Tanner, K., Eggett, D., & Corbn-Lewis, K. (2020). Age-related changes in speech and voice. Spectral and Cepstral Measures. *JSLHR* 63 (3), 647-660. https://doi.org/10.1044/2019_JSLHR-19-000.

Tursunov, A.; Mustaqeem; Choeh, J.Y.; Kwon, S. Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms. *Sensors* 2021, *21*, 5892.

Acoustic and kinematic correlates of adaptive responses to consistent formant perturbations in young healthy speakers

Teja Rebernik[1,2], Tomas O. Lentz[3,4], Hayo Terband[4,5]

[1]University of Groningen
[2]Vrije Universiteit Brussel
[3]Tilburg University
[4]Utrecht University
[5]University of Iowa

t.rebernik@rug.nl, t.o.lentz@tilburguniversity.edu, hayo-terband@uiowa.edu

Introduction. Auditory feedback perturbation (AAF) paradigms are a common way to investigate speech motor control. One type of AAF tasks, established by Houde and Jordan (1998), are adaptive formant perturbation tasks, where vowel formants are gradually changed in real time in order to assess individuals' ability to integrate auditory feedback and change their speech production after a perceived mismatch between what they expected to hear versus what they actually heard. Prior studies, which predominantly investigate adaptive responses captured acoustically, show that participants tend to respond to gradual upward formant perturbations by shifting their own formant production downward (e.g., Nault & Munhall 2020; Villacorta et al. 2007). While responses to formant perturbation tasks are fairly robust, the adaptation is not complete and, additionally, not all participants adapt to the same extent. One potential reason could be that even though the auditory feedback is changed, somatosensory feedback stays the same: speakers' adaptation patterns might therefore vary depending on their individual relative weighing of importance of auditory versus somatosensory feedback (e.g., Katseff et al. 2012; Lametti et al. 2012).
In this study, we investigate both acoustic as well as kinematic correlates of AAF responses, with the aim of determining whether kinematic data can help explain variability captured in the acoustic responses. In line with prior studies examining the relationship between tongue positions and formant frequencies, we expect tongue and jaw height to be inversely related to F_1 (e.g., Lee et al. 2016). In addition, we expect that the measure in which the potential adaptive response is the largest is not the same across all participants.

Methods. The data presented in the abstract are part of a larger study, which was approved by the Ethics Assessment Committee of Utrecht Institute of Linguistics - OTS at Utrecht University. A total of seven speakers (five females, two males; mean age: 21.3 years, SD: 3 years) are analysed here (the full dataset comprises 41 speakers). The participants were instructed to repeat three Dutch words ('peer' /pɪːr/, 'beer' /bɪːr/ and 'veer' /fɪːr/, meaning 'pear', 'bear', and 'feather', respectively) while their F_1 and F_2 were gradually shifted using Audapter (Cai et al. 2012). After 27 trials of veridical feedback ('baseline' phase), their formants were gradually perturbed for 24 trials ('ramp' phase) until reaching the maximum perturbation of a 25% increase in F_1 and 12.5% decrease in F_2 ('hold' phase; perturbation towards /a/). Following 27 trials in the 'hold' phase, the perturbation was suddenly dropped for the last 24 trials ('release' phase). For more details on the paradigm, see Van Brenk and Terband (2020). The data was synchronously collected both acoustically, using an over-the-ear microphone (T-bone EM 9600), and kinematically, using AG501 electromagnetic articulography sensors. Besides reference sensors, which were placed on the upper incisor and the left and right mastoid processes, movement sensors were placed on the tongue (tip, body and dorsum), the lips (vermillion border of the upper and lower lip, left and right oral commissures), and the jaw (lower incisor).
After annotating the participants' productions, we coupled acoustic and head-corrected kinematic data using an in-house R script. We calculated the mean F_1 and F_2 frequency (in Hz) as well as mean tongue and jaw height coordinates (TT_y, TB_y, TD_y, Jaw_y; in mm) between 40-120ms of each production. The data was then normalized to the *baseline* phase, allowing us to be able to assess changes in acoustics and kinematics in the *hold* phase relative to the *baseline*. To statistically analyze our data, we used linear mixed-effects models, as implemented in the *lme4* package (Bates et al., 2015) in R Studio (R version 4.3.1). Our model included normalized values as the dependent variable, phase (levels: *baseline*, *hold*) in interaction with measure type (levels: F_1, F_2, TT_y, TB_y, TD_y, Jaw_y) as the main fixed effect, and word as well as sex as additional fixed effects. We included a by-participant random intercept and phase as a by-participant random slope to account for differences across speakers. Including as a by-participant random slope was not supported by the data and resulted in a model's failure to converge. Confidence intervals were calculated using the *confint* function of the *lme4* package.

Results. At the group level in the acoustic domain, there was no evidence for an effect of phase on F_1 ($\beta = -0.04$, $p = 0.9$, 95% CI = [-0.53, 0.45]) but there was an effect on F_2 ($\beta = 0.5$, $p = 0.049$, 95% CI = [0.04, 1.03]), with participants increasing their F_2 in response to the perturbation. In the kinematic domain, there was a significant group-level effect of

phase on TT_y (β = -0.6, p = 0.03, 95% CI = [-1.09, -0.03]), TD_y (β = -0.8, p = 0.003, 95% CI = [-1.36, -0.26]) and Jaw_y (β = -0.6, p = 0.04, 95% CI = [-1.11, -0.08]) but not on TB_y (β = -0.5, p = 0.07, 95% CI = [-0.99, 0.01]). In addition to the statistical analysis, we also visually evaluated individual speaker patterns, which showed that the participants were variable in their behavioral response to the perturbation. Some speakers adapted multiple acoustic and kinematic parameters while others adapted only one or even none, and speakers differed in the measure where their adaptation was most apparent. Furthermore, some speakers showed clear corresponding changes in both acoustic and articulatory space, while other speakers did not. One such example of a speaker who did not significantly change their formant production but did change their articulatory movements is displayed in Figure 1. While the produced formants were not apparently changed in the *hold* compared to the baseline phase (as evidenced by a large degree of overlap between the two phases), tongue body and dorsum *were* higher (which would be the expected adaptation pattern for an upwards F_1 perturbation, considering the inverse tongue height – F_1 relationship) while tongue tip and jaw were lower. Detailed results on individual patterns will be available at the conference.

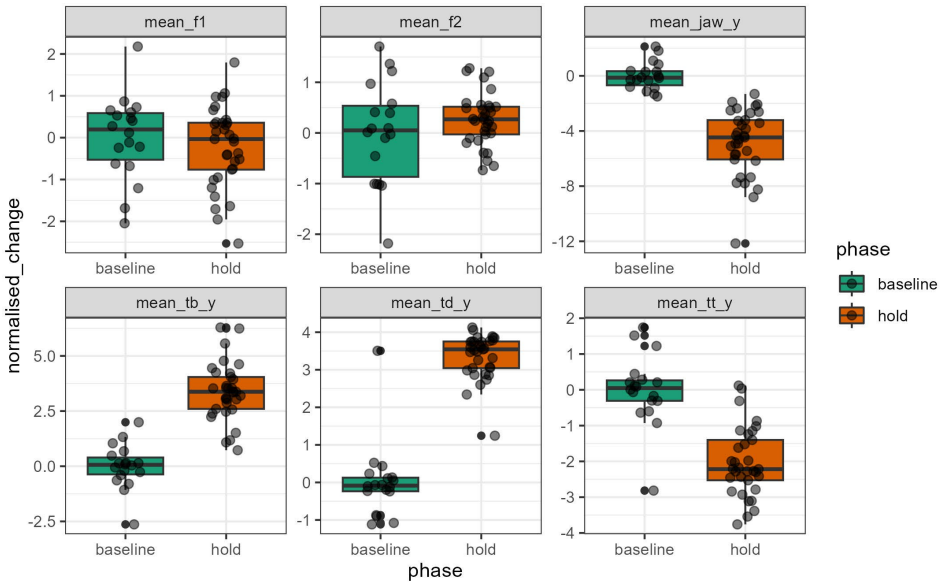


Figure 1: *Example of a male speaker who did not adapt acoustically but did adapt in his kinematics, by raising his tongue body and dorsum and lowering his jaw and tongue tip. Note the different scales on the y-axis, however, used for clarity of presentation.*

Discussion. Current results indicate that there might be some adaptation patterns that are seen in kinematics but not acoustics (namely F_1). The study does face some limitations, including a limited set of speakers with an imbalanced sex distribution. Consequently, while some individual patterns for kinematic versus acoustic correlates of adaptive responses were visible through a visual examination, this could not yet be statistically confirmed. Further analysis will therefore include the full dataset (we have collected data from 41 speakers), as well as a detailed exploratory analysis. The study so far, however, does show that the kinematic correlates behind adaptive responses to formant perturbations could potentially partly explain the variability in responses seen in the literature.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). "Fitting Linear Mixed-Effects Models Using lme4." In: *Journal of Statistical Software*, 67.1, pp. 1–48.

Cai, S., Beal, D. S., Ghosh, S. S., Tiede, M. K., Guenther, F. H., & Perkell, J. S. (2012). "Weak Responses to Auditory Feedback Perturbation during Articulation in Persons Who Stutter: Evidence for Abnormal Auditory-Motor Transformation." In: *PLoS One*, 7.7, pp. e41830.

Houde, J., & Jordan, M. I. (1998). "Sensorimotor adaptation in speech production." In: *Science Reports*, 279.5354, pp. 1213–1216.

Katseff, S., Houde, J., & Johnson, K. (2012). "Partial Compensation for Altered Auditory Feedback: A Tradeoff with Somatosensory Feedback?" In: *Language and Speech*, 55.2, pp. 295–308.

Lametti, D., Nasir, S. M., & Ostry, D. J. (2012). "Sensory Preference in Speech Production Revealed by Simultaneous Alteration of Auditory and Somatosensory Feedback." In: *The Journal of Neuroscience*, 37.27, pp. 9351–9358.

Lee, J., Shaiman, S., & Weismer, G. (2016). "Relationship between tongue positions and formant frequencies in female speakers." In: *The Journal of the Acoustical Society of America*, 139.6, pp. 426–440.

Nault, D. R., & Munhall, K. G. (2020). " Individual variability in auditory feedback processing: Responses to real-time formant perturbations and their relation to perceptual acuity". In: *The Journal of the Acoustical Society of America*, 148.6, pp. 3709–3721.

Van Brenk, F., & Terband, H. (2020). "Compensatory and adaptive responses to real-time formant shifts in adults and children." In: *The Journal of the Acoustical Society of America*, 147.4, pp. 2261–2270.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). "Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. In: *The Journal of the Acoustical Society of America*, 122.4, pp. 2306–2319.

An LSTM analysis of timescales of breakdown in ALS

Jessica Campbell[1], Louis Goldstein[1], Khalil Iskarous[1]

[1]University of Southern California

jac95339@usc.edu, louisgol@usc.edu, kiskarou@usc.edu

Introduction. Artificial Intelligence (AI) techniques are increasingly applied to the clinical realm, employing machine learning methods to diagnose speech disorders or assist patients (e.g., Millet & Zeghidour 2019, Chandrakala & Rajeswari 2017). Since motor control-based disorders lead to complex, difficult to diagnose speech patterns, deep learning could be ideal for classification as hidden neurons can encode abstract features, making solution of the typical/atypical classification task feasible despite nonlinearity and nonconvexity. The current study applies machine learning methods to articulatory data to determine the loci of breakdown during articulation in patients with amyotrophic lateral sclerosis (ALS). This disease is associated with breakdown at two different linguistic hierarchical levels that occur at gestural and stress foot timescales ("gestural" and "prosodic" levels, respectively) (e.g., Weismer et al. 2003, Kim et al. 2009). At the gestural level, ALS speech has shown hints of temporal discoordination in the form of excessive variability during production. However, highly varied methods and research questions have resulted in general uncertainty about the diagnosis of breakdown at this linguistic level (e.g., Romö, Lee, & Robb 2021, Weismer et al. 2003, Yunusova et al. 2008, Kuruvilla et al. 2012). New methods are employed in the current study in an effort to diagnose breakdown at this level; we hypothesize that breakdown is indeed present at the gestural level, as some research has found. At the prosodic level, previous research has generally reached a consensus that ALS speech in English exhibits less distinction between stressed and unstressed syllables than typical speech does (Liss et al. 2009, Kim et al. 2021). Following previous research, we hypothesize that breakdown is present in ALS speech at the prosodic level, as well. There are thus two different issues of focus in ALS speech at two different levels of the linguistic hierarchy. Since these two linguistic levels exist at different temporal scales, our technique probes the difference through control of window size (memory) of the neural network.

Methods. Speech of control and ALS participants was sourced from the X-Ray Microbeam Corpus (Westbury, Turner, & Dembowsky 1994) and one additional study using the same methods (Weismer et al. 2003), collected at the University of Wisconsin, Madison. Data consisted of articulatory pellet-tracking data with stimuli of read syllables, sentences, and paragraphs at comfortable tempos from 18 speakers (6 ALS, 12 control), totaling 19 minutes, 51 seconds of ALS speech and 19 minutes, 11 seconds of control speech. Position data was narrowed to only relevant articulators and axes; it thus comprised lip aperture, ventral tongue in the horizontal direction, ventral tongue in the vertical direction, mid-tongue in the vertical direction, and jaw in the vertical direction. Lip aperture was calculated as the Euclidean distance between the X and Y coordinates of the upper lip and lower lip at every time sample. Velocity was calculated from the coordinates of all markers excluding the upper and lower lip by calculating the Euclidean distance between each successive marker position. For lip aperture, velocity was calculated as the absolute value of change across successive lip aperture values.
A bidirectional long short-term memory network (bi-LSTM) was implemented using Keras (Chollet et al. 2015) to classify both types of time-varying data into ALS and control groups. LSTMs are ideal, since they apply to time series, and do not require extraction of landmark measures. The bi-LSTM then fed into a fully-connected five-neuron layer, then into a softmax classifying layer that determined which group the section of data had the highest probability of belonging to. Two LSTM parameters were varied in the study: window size and LSTM complexity. Window size determines the maximum length of memory the LSTM can access at once and thus allows us to determine the time scale at which the temporal breakdowns of interest occur. A small window size would limit memory so that fluctuations or patterns over long timescales would not be able to be learned by the network. For example, window sizes much smaller than a stress foot would not learn to classify data based on lack of contrast between stressed and unstressed syllables. For the gestural level, the window size was 40ms (6 frames), slightly shorter than a single gesture. At the prosodic level, it was 404ms (59 frames), as close as possible to the average stress foot duration from one acoustic stressed vowel onset to the next (Campbell et al. 2023). Complexity was measured by the number of LSTM neurons (neurons with forget gates) in the hidden layer and directly relates to the level of abstraction of the features being extracted to classify the data; increasing complexity can lead to the resolution of increasingly abstract and complex aspects of the data. Complexity was varied from 2 to 30 hidden neurons in steps of 2. Four seeds were used to randomize the data order, with five runs in each seed for each level of complexity and window size. There were therefore 600 runs in total for each data type. Percent accuracy was calculated as the number of windows in the test set (13-16% of total frames depending on the run) that were properly classified. We predicted that the LSTM would more easily classify data in which the ALS group showed behavior unlike the control speakers. Given our hypotheses that breakdown in ALS speech occurs at both the gestural and prosodic levels, we predicted higher than chance classification accuracy both when the LSTM was provided with a small window size

that encapsulated gestural level information and when it was provided with a large window size that encapsulated prosodic level information.

Results and discussion.

Our first hypothesis is that there is breakdown in ALS speech at the gestural level, and this predicts that an LSTM would be able to classify the two groups with higher than chance accuracy when window size (i.e., memory) is small and only captures gestural-level timing patterns. The data supports this hypothesis: for all data types, the small window condition led to accuracy above chance (77.5% [SD = 3.28] for position data, 61.46% [SD = 4.45] for velocity data). We further hypothesized that, due to breakdown at the prosodic level in ALS speech, the LSTM would be able to successfully classify the groups when window size was large (i.e., the size of a stress foot). This prediction was also observed for all data types, with accuracy above chance for the large window size (82.20% [SD = 7.24] for position data, 67.30% [SD = 6.25] for velocity data). A linear mixed model for each data type predicting accuracy from window size, log-transformed LSTM units, and their interaction with random intercepts grouped by seed revealed higher accuracy for the large window size than the small window size (p<0.001 for both data types). Increasing log-transformed LSTM complexity significantly increased accuracy for both conditions (p < 0.001 for both data types), though it increased it more steeply for the large window size (p<0.001 for both window types); this may be attributed to the fewer ways that the hidden layers can analyze the data in the smaller window size condition. In conclusion, we have found evidence of breakdown in ALS speech at two different timescales, using machine learning parameters to examine specific linguistic hierarchical levels. Further probing of the hidden layer in the future may indicate the types of breakdown occurring at these levels.

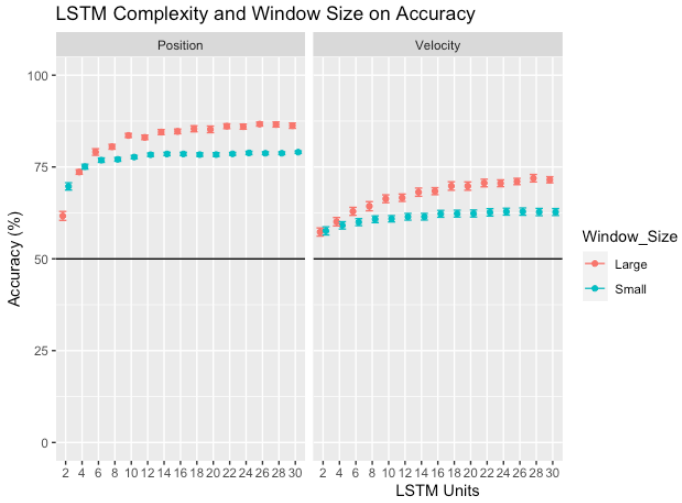


Figure 1: *Accuracy of classification based on data type and LSTM parameters window size and complexity (number of hidden units in the hidden layer). Error bars represent standard error.*

References

Campbell, J., Byrd, D., Goldstein, L. (August, 2023). Viable signal periodicities in speech rhythm. In Radek Skarnitzl & Jan Volín (Eds.), Proceedings of the 20th International Congress of Phonetic Sciences. Guarant International. 659-663.

Chandrakala, S., & Rajeswari, N. (2016). Representation learning based speech assistive system for persons with dysarthria. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(9), 1510-1517.

Chollet, F. and others (2015). Keras. https://keras.io

Kim, D., Kuruvilla-Dugdale, M., de Riesthal, M., Jones, R., Bagnato, F., & Mefferd, A. (2021). Articulatory correlates of stress pattern disturbances in talkers with dysarthria. *Journal of Speech, Language, and Hearing Research*, *64*(6S), 2287-2300.

Kuruvilla, M. S., Green, J. R., Yunusova, Y., & Hanford, K. (2012). Spatiotemporal coupling of the tongue in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 55*(6), 1897-1909.

Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., & Caviness, J. N. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech, Language, and Hearing Research, 52*(5), 1334-1352.

Millet, J., & Zeghidour, N. (2019, May). Learning to detect dysarthria from raw speech. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Proceedings. IEEE. 5831-5835.

Romö, N., Lee, J., & Robb, M. P. (2022). Properties of relative timing and phonetic complexity in adults with dysarthria secondary to amyotrophic lateral sclerosis. *Folia Phoniatrica et Logopaedica*, *74*(4), 284-295.

Weismer, G., Yunusova, Y., & Westbury, J. R. (2003). Interarticulator coordination in dysarthria. *Journal of Speech, Language, and Hearing Research, 46*(5), 1247-1261.

Westbury, J. R., Turner, G., Dembowski, J. (1994.) X-ray Microbeam Speech Production Database User's Handbook. University of Wisconsin.

Yunusova, Y., Weismer, G., Westbury, J. R., & Lindstrom, M. J. (2008). Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research, 51*(3), 596-61.

Exploring F0 Entrainment in Bi-directional Speech Imitation: A Cross-Linguistic Analysis

Zheng Yuan[1,2], Štefan Beňuš[3,4], Alessandro D'Ausilio[1,2]

[1]*Italian Institute of Technology, Italy*
[2]*University of Ferrara, Italy*
[3]*Constantine the Philosopher University, Slovakia*
[4]*Slovak Academy of Sciences, Slovakia*
{zheng.yuan, Alessandro.Dausilio}@iit.it, sbenus@ukf.sk

Introduction.
Speech Entrainment (Wynn and Borrie 2022) is a complex phenomenon that refers to the ways in which speakers adapt their speech to align with the acoustic-prosodic characteristics of their conversational partners. Studies on the neuro-psycho underpinnings of speech entrainment have revealed that it can influence the development of second language (L2) pronunciation and prosody as a way of implicit speech imitation (Delvaux and Soquet 2007).
Aligning with Wilt et al. (2023) that "automatic imitation is enhanced for non-native sound", this study aims to address a research gap in L2-L2 interaction by investigating the degree of F0 entrainment in dyadic L2 English speech imitation among speakers whose L1 background is Italian, French, or Slovak. Specifically, we aim to examine the relationship between L2 English proficiency and F0 entrainment, while considering other variables such as sentence length, interaction time, and intra-dyad language score differences.

Methods.
We use the alternating reading task (ART, see Yuan et al. (2023)) dataset for investigating phonetic convergence in the context of L2 acquisition. The dataset comprises recordings of 58 participants who were native Italian, French, and Slovak speakers. In the ART experiment, speakers make dyads and take turns reading aloud a neutral English text that contained 80 sentences with turn boundaries set within sentences. The experiment involved three experimental conditions: solo, interactive, and imitation.
To assess the relationship between F0 entrainment and L2 proficiency, we evaluated the spoken English scores of the participants in the solo reading condition. Six language experts, three of whom were native Chinese speakers and three of whom were native Slovak speakers, evaluated the participants' spoken English using four scoring indicators: pronunciation, intonation, fluency, and overall impression. For each indicator, the evaluators scored the participants on a scale from 1 to 5. The defined spoken English score for each speaker is the average sum of the 4 indicators over the evaluators. The degree of agreement among these experts for each criterion is quantified by the Intraclass Correlation Coefficients (ICC) values with 95% confidence intervals. Notably, the overall ICC value (0.657) demonstrates a statistically significant (p-value < 0.001) and moderately reliable level of agreement.
In this study, F0 was extracted using autocorrelation in PRAAT software with a sample rate of 100 Hz. To ensure accurate and smooth F0 contours, voiceless utterances were removed, and F0 outliers were bridged by linear interpolation. A two-pass method (De Looze and Rauzy 2009) was also applied to handle F0 outliers. Following this, a Savitzky-Golay filter was used to smooth the F0 contour with third-order polynomials in 7-sample windows. Finally, the F0 contour was parameterized using four features, including slope, mean, range, and drop, to provide a quantitative analysis of the F0 contour's shape (Reichel, Beňuš, and Mády 2018).
We first extracted an ASR-generated transcription with word-level time alignment, which was obtained using WhisperX (Bain et al. 2023). To measure the degree of entrainment, the distance between two parameterized pitch contours was calculated. The dynamic time warping algorithm was used with Euclidean distance as a similarity measure to align the pitch contours.

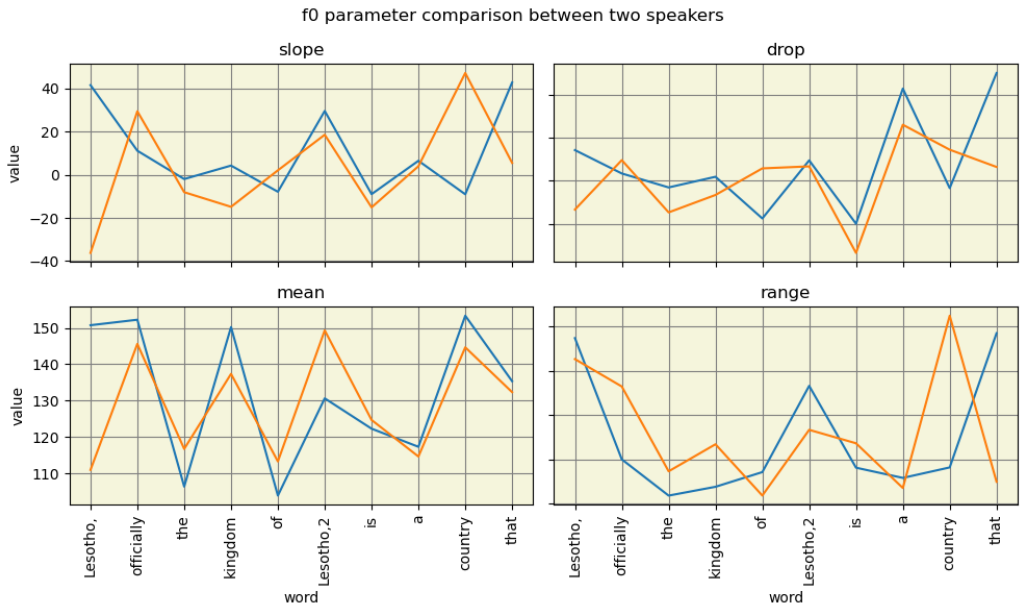


Figure 1: Pitch Parameterisation

Preliminary Results.
The intra-dyad Spoken English Score difference for Slovak speakers has a significant correlation with the intra-dyad distance of mean F0 contours (Pearson's r = -0.36).

Discussion. Our preliminary results have partially support the hypothesis that speakers with higher spoken English proficiency tend to have more accurate imitation of their interlocutors' pitch contours. This finding aligns with Yuan et al. (2023) where the imitation ability was measured by the LexTale score (Lemhöfer and Broersma 2012) instead of the speaking skills. In the full paper, we plan to extend the experiment to the other three F0 parameters and conduct comparative analysis across the Italian, French, and Slovak speakers.

References.

Bain, Max, Jaesung Huh, Tengda Han, and Andrew Zisserman (2023). "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio". In: *INTERSPEECH*.

De Looze, Céline and Stéphane Rauzy (2009). "Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration". In: *Interspeech 2009*.

Delvaux, Véronique and Alain Soquet (2007). "The influence of ambient speech on adult speech productions through unintentional imitation". In: *Phonetica* 64.2-3, pp. 145–173.

Lemhöfer, Kristin and Mirjam Broersma (2012). "Introducing LexTALE: A quick and valid lexical test for advanced learners of English". In: *Behavior research methods* 44, pp. 325–343.

Reichel, Uwe D, Štefan Beňuš, and Katalin Mády (2018). "Entrainment profiles: Comparison by gender, role, and feature set". In: *Speech Communication* 100, pp. 46–57.

Wilt, Hannah, Yuchunzi Wu, Bronwen G Evans, and Patti Adank (2023). "Automatic imitation of speech is enhanced for non-native sounds". In: *Psychonomic Bulletin & Review*, pp. 1–17.

Wynn, Camille J and Stephanie A Borrie (2022). "Classifying conversational entrainment of speech behavior: An expanded framework and review". In: *Journal of Phonetics* 94, p. 101173.

Yuan, Zheng, Aldo Pastore, Dorina de Jong, Hao Xu, Luciano Fadiga, and Alessandro D'Ausilio (2023). "The ART of Conversation: Measuring Phonetic Convergence and Deliberate Imitation in L2-Speech with a Siamese RNN". In: *Proc. INTERSPEECH 2023*, pp. 132–136. DOI: 10.21437/Interspeech.2023-2283.

Insights into phonemes' articulation time

Montse Soberanes[1], Carlos A. Pérez-Ramírez[2], M. Florencia Assaneo[1]

[1]*Universidad Nacional Autónoma de México*

[2]*Universidad Autónoma de Querétaro, México*

montsesm13@gmail.com, carlos.perez@uaq.mx, fassaneo@inb.unam.mx

Introduction. Speech production is a dynamic process, involving the careful cooperation between articulators (i.e, tongue, jaw, lips and velum) and vocal folds to produce the elemental sounds that create words, referred to as phonemes; Twaddell (1935). Phonemes have long been studied, and nowadays much is known about their acoustic properties and the articulatory configurations required for their production, e.g., Stevens (2005). Nevertheless, phonemes have temporal aspects that remain unexplored. The time it takes to produce a phoneme is highly variable, even in the same language. For example, in Spanish, this time ranges from 30 to 150 ms (Barrio & Torner 1999; Marín 1995). Even though we can find such an ample range, little has been dedicated to understanding its origin. Trying to fill this gap in knowledge, the current work explores 3 plausible factors modulating phonemes' articulation time: attention, coarticulation and fast/slow intended speech speed.

Methods. Participants (n=20) connected to an electromyography system (EMG) to record lips muscle (*orbicularis oris*) activity and placed close to a microphone to record their vocalisations, completed 4 articulation blocks. On each block, they were instructed to pronounce a syllable (/pɑ/ or /pu/) right after hearing a tone (120 cue tones were included per block, with a random inter stimulus interval of 0.75 to 3.6 s). Each articulation block was preceded by a speed priming step, where participants listened to a rhythmic train of tones while concurrently and repeatedly whispering the syllable /pe/, trying to match the external rhythm. Two priming speeds were tested: 3 and 5 syll/s. Additionally, participants' attentional state was assessed by means of a classic Flanker task at the beginning, middle and end of the whole protocol. Level of attention was assigned to each articulation block as the percentage of correct responses of the nearest Flanker task.

The articulation time for the /p/ was computed as the difference between the speech onset (i.e., the burst sound corresponding to the release of the occlusion, obtained from the acoustic signal) and the onset of the lips muscle activity (i.e., the beginning of the motor gesture, extracted from the EMG recordings). As for the vowels, the articulation time was determined by the estimated voiced time obtained from the acoustic signal. Two linear mixed effect model analyses were performed, one to predict the duration of the /p/ and another one for the duration of the vowels. In both cases, a backward elimination was performed starting from a model including: priming speed, attentional state and vowel (consequent vowel for the /p/ and phoneme identity for the vowels) as fixed factors. Intercepts, but not slopes, were allowed to vary per participant. The models that better explained durations were chosen based on the change in Bayesian Information Criterion (BIC).

Results. The model that better predicted the duration of the /p/ included attentional state and consequent vowel, but not priming speed. Accordingly, we computed the estimated marginal means for the factors included in the model. We found a positive linear relationship between /p/ duration and attentional state (trend=2.65, $p<0.001$; see Fig. 1A) and shorter duration times when the consequent vowel is /ɑ/ rather than when the consequent vowel is /u/ ($mean_u$=192 ms, $mean_ɑ$=176 ms, $p<0.001$; see Fig. 1B).

As seen for /p/, the model that better adjusted vowels' duration (/ɑ/ and /u/ in this case) included attentional state. However, in contrast with the previous result, this model comprised priming speed but left out phoneme identity. As in the /p/, higher attention levels led to longer phoneme durations (trend=1.48, $p=0.0016$; see Fig. 1C). Additionally, we

found that conditions primed at 5 syll/s gave place to shorter vowels than the ones primed at 3 syll/s ($mean_3$=304 ms, $mean_5$=299 ms, p<0.001; see Fig. 1D).

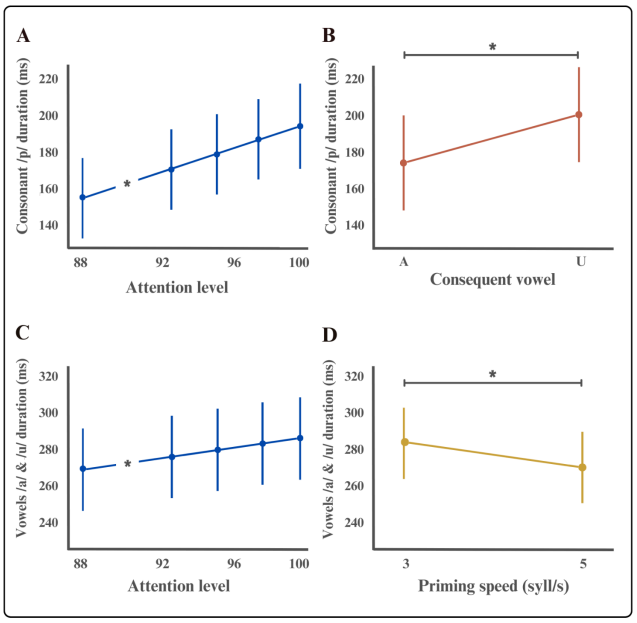


Figure 1: *Linear mixed effect models results. A&B. Predicted /p/ duration as a function of attention level and consequent vowel, respectively. C&D. Predicted vowels duration (/a/ and /u/) as a function of attention level and priming speed, respectively. Dots: model predicted group means. Bars: 95% confidence interval. * p < 0.002*

Discussion. We observed attentional status significantly modulated duration across all the analysed phonemes, with higher attention levels resulting in longer production times. When attention level is high the articulation of every phoneme is carefully done producing a higher quality speech, Dromey and Shim (2008), this aligns well with the observed positive relationship between phoneme's duration and attentional state. The consonant /p/ is additionally affected by the consequent vowel, which can be explained by the coarticulation phenomenon. Rounded lips are required to produce the /u/ but not the /ɑ/. This feature may be inherited by the /p/ resulting in longer times when it is followed by /u/ rather than when it is followed by /ɑ/. Surprisingly, the vowels' duration are not affected by phoneme identity, meaning that /ɑ/ and /u/ don't have significantly different durations. However, they are affected by the priming speed, with longer times for slower priming rates. The fact that priming speed affects vowels but not consonants suggests that when speaking faster or slower, the phonemes adapting their durations are the vowels, while consonants remain unchanged, which falls in line with Fujimura's (1981) observation of consonantal gestures being rigid and not subject to speed modulation and is consistent with theories proposing consonants as intermediate dynamical states connecting vowels; Browman & Goldstein (1989).

References.

Barrio Estévez, L. D., & Torner Castells, S. (1999). La duración consonántica en castellano. In: *ELUA. Estudios de lingüística*, 13, pp 9-35.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. In: *Phonology*, 6(2), 201-251.

Dromey, C., & Shim, E. (2008). The effects of divided attention on speech motor, verbal fluency, and manual task performance. In: *American Speech-Language-Hearing Association;* 5(1), pp 1171–1182.

Fujimura, O. (1981). Temporal Organization of Articulatory Movements as a Multidimensional Phrasal Structure. *Phonetica*, 38, 66-83.

Marín Gálvez, R. (1995). La duración vocálica en español. In: *ELUA. Estudios de Lingüística,* 10, pp. 213-226.

Stevens, K. N. (2005). The acoustic/articulatory interface. In: *Acoustical science and technology*, 26(5), pp 410-417.

Twaddell, W. F. (1935). On Defining the Phoneme. In: *Language,* 11(1), 5–62. https://doi.org/10.2307/522070

Exploration and classification of vocal fry, period doubling, and modal voice using acoustic and EGG measures

Yaqian Huang

Acoustics Research Institute, Austrian Academy of Sciences

yaqian.huang@oeaw.ac.at

Introduction. While it is generally agreed upon that creaky voice and modal voice differ in their acoustic and articulatory properties, it is less clear how subtypes of creaky voice differ among each other in those aspects. Common acoustic attributes of creaky voice include low f0, low spectral tilt, and noise (Garellek 2019), which are expected to various extents for subtypes of creaky voice. For example, according to Keating *et al*. 2015, vocal fry is typically characterized as having low f0 and spectral tilt, and damped pulses. Period doubling, in contrast, contains two alternating glottal cycles which differ in amplitude or frequency (Kreiman *et al*. 1993), and typically has noise, low spectral tilt, and high subharmonics. The distinctions between subtypes of creaky voice have been noted and substantiated in several classification schemes, but mainly manually based on their acoustic waveforms (Hedelin & Huber 1990; Redi & Shattuck-Hufnagel 2001). There lacks a systematic assessment of the importance of both acoustic and articulatory measures, given that the voice source defines the major dimension of phonation differences.

This study contributes by including electroglottographic (EGG) measures (that are used to quantify the degree of vocal fold contact) in addition to acoustic measures to assess the importance of both source and filter characteristics to differences between vocal fry and period doubling. Two machine learning algorithms were used to evaluate the effects of acoustic and articulatory measures on the classification of subtypes of creaky voice. Multinomial logistic regression with *l1* regularization (Lasso) was used to test the classification performance using these measures as feature representation of creaky voice. A separate random forest model was used to examine the feature importance of each measure. This study uses continuous read speech in Mandarin from multiple speakers, as vocal fry and period doubling were commonly found allophonic to Mandarin tones (Yu, 2010). The findings clarify the similarities and differences within creaky voice and between creaky and modal voice, and will be of practical interest to speech and voice detection.

Methods. The EGG and audio corpus consists of read speech recorded from 20 native Mandarin speakers (10 F; mean age: 20.1; range = 18-22) (Huang 2022). The fixed carrier sentences embedded varying trisyllabic words: *wo3 tɕau1 niˇ3 WORD tsən3-mɤ0 ʂʷo1* "I teach you WORD how to say". Picture fillers were used every four sentences, and participants were asked to briefly describe the object that the picture showed. Each recording session contained 480 sentences and lasted about 45 minutes. The EGG recordings were band-pass filtered between 40 and 22050 Hz with smoothing at 50 Hz in Praat to remove the low-frequency DC component and higher frequency noise. 638 tokens of vocal fry and 3297 tokens of period doubling were identified using the EGG visually based on canonical characteristics (see Kreiman *et al*. 1993; Keating *et al*. 2015); non-vocalic segments were verified in the audio waveforms and excluded. 1603 tokens of modal voice were sampled from adjacent regions of period doubling or vocal fry based on EGG and verified in audio.

For each token, mean acoustic and EGG measures were extracted using VoiceSauce (Shue *et al*. 2011) and EGGWorks (Tehrani 2009), respectively. Praat's f0 algorithm adjusted to detect the longest period (a pair of alternating cycles) in period doubling was used as the basis of further acoustic measures (e.g., spectral tilt). Incorrect f0s were checked and excluded based on each speaker's pitch range calculated from EGG. 32 acoustic measures included corrected harmonics and spectral tilts, harmonics-to-noise ratios, subharmonic-to-harmonic ratio, formants and bandwidths, and energy measures; 11 EGG measures included contacting and decontacting durations, contact quotient, speed quotient, and cycle peak velocity measures. All measures were scaled to a standard normal distribution.

Because of the exploratory nature of the analysis among three voice types, I first used t-distributed stochastic neighbor embedding (t-SNE), a dimensionality reduction technique to compare the similarity among all tokens in high-dimensional datasets (van der Maaten & Hinton 2008). Given the correlations within a particular family of acoustic/articulatory measures and between different families of measures, I then used logistic Lasso regression to shrink coefficients of less informative predictors, which helps reduce multicollinearity and overfitting issues and enables variable selection. I also used a random forest model to classify and predict these voicing types and compare the results with those of the logistic regression. Two datasets were used: acoustic measures (5538 rows x 33 cols) or a combination of acoustic and EGG measures (918 rows x 44 cols; data were sparser for tokens which have shared acoustic and EGG measures); a binary-coded gender factor was added to the predictors. Cross validation was used by splitting the dataset into a training set (~66.7%) and a test set (~33.3%). The training and the test sets had similar distributions of the different voicing types. All models were first devised using the training set, and then evaluated in the test set.

Results. T-SNE clusters based on voicing types are shown in **Figure 1**. Both as creaky voice, period doubling and vocal fry are spatially closer than modal voice whereas period doubling appears to be closer to modal voice than vocal fry does. Adding EGG measures helps separate the subtypes of creaky voice.

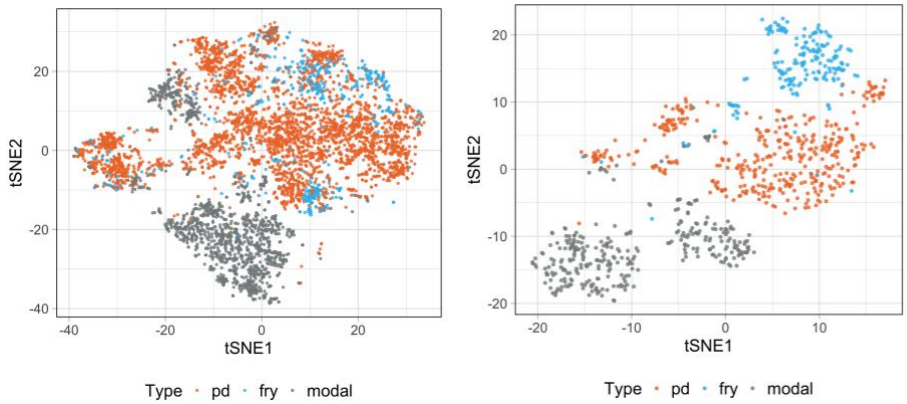


Figure 1: *t-SNE clustering of tokens of period doubling (orange), vocal fry (blue), and modal voice (gray) using only the acoustic (left) and both acoustic and articulatory measures (right). The datasets have different sizes.*

Table 1 shows the results of the model performance in the test set of the two datasets using acoustic and/or EGG measures. The results suggest that both models achieved comparable performance. With only acoustic features, both models were able to achieve decent recall at around 85%. Adding EGG measures improved precision and recall in both models substantively. This indicates that articulatory features provide crucial information in distinguishing among voice qualities.

Table 1: *Summary of overall accuracy and macro average precision and recall scores using different machine learning methods.*

	ML approach	Accuracy	Macro avg. precision	Macro avg. recall
Acoustics	Logistic Lasso regression	0.9112	0.8749	0.8137
	Random forest	0.9312	0.9098	0.8529
Acoustics + EGG	Logistic Lasso regression	0.9837	0.9840	0.9784
	Random forest	0.9967	0.9975	0.9970

With the feature elimination given by the logistic Lasso regression, the non-zero coefficients signal the most distinctive predictors that contribute to a certain voice category. In the dataset of acoustic and EGG measures, more predictors were shrunk to zero, suggesting that the addition of the phonatory dimension improves the model and makes it more interpretable with fewer distinctive features. For example, modal voice is captured by higher H1*, H1*–H2*, and f0, vocal fry is captured by lower H4*, higher contact quotient and cycle minimum velocity, and lower speed quotient, and period doubling is captured by higher H4* and H4*–2K*, lower H1*–H2*, and lower decontacting duration.
Further, the random forest model ranks the variables according to their importance based on classification accuracy and Gini index (a node-based tree evaluation metric). Among all the acoustic measures, f0, H1*–H2*, H1*, SoE, H2*, and HNR<500Hz are the most important; further, the decontacting duration and contact quotient are the most important EGG measures among all.

Discussion. The models using acoustic features alone already show reasonable separation whereas the models using a larger set with both acoustic and articulatory features effectively distinguish period doubling, vocal fry, and modal voice from each other. The most important acoustic measures are f0, H1*–H2*, and H1*, and the most important EGG features are the duration of the glottal opening phase and contact quotient, as established by drawing from both random forest models and logistic Lasso regression. It is not surprising, though, that the voice qualities are better captured when EGG measures are added in the models, especially for the two subtypes of creaky voice. Voicing types have stronger ties to the source dynamics associated with our vocal folds, and could appear acoustically similar and are better distinguished by phonatory measures.
However, considering the mapping between perception and acoustics, phonatory measures are hardly accessible to listeners when encountering speech signals. Though adding the phonatory dimension better differentiates subtypes of creaky voice and modal voice in production, in perception, it remains unclear whether and how phonatory characteristics are transmitted to influence people's perception. It implies that listeners may show less robust categorization choices than a machine does with all the available acoustic and articulatory features in speech and voice detection.

References. Garellek, M. (2019). The phonetics of voice. Hedelin, P. and Huber, D. (1990). Pitch period determination of aperiodic speech signals. Huang, Y. (2022). Articulatory properties of period-doubled voice in Mandarin. Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. Kreiman, J., Gerratt, B. R., Precoda, K., and Berke, G. S. (1993). Perception of supraperiodic voices. Redi, L. and Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: Tehrani, H. (2009). EGGWorks. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE.

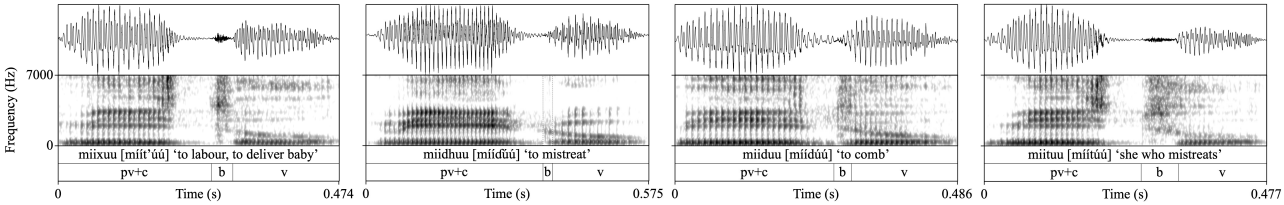
Perception of a four-way stop laryngeal contrast in Eastern Oromo

Maida Percival[1]

[1]University of Toronto
maida.percival@mail.utoronto.ca

Introduction. Eastern Oromo is an East Cushitic language of Ethiopia, and is uncommon among the world's languages in having a four-way stop laryngeal contrast that includes an ejective and implosive stop at the same (coronal) place of articulation. The present study examines the perceptual cues to this laryngeal stop contrast, where an ejective, implosive, voiced pulmonic, and voiceless pulmonic stop contrast. Previous work has examined the production of stop contrasts in the language (Percival 2014; Percival, Kochetov, and Kang 2018) as well as perception of Oromo listeners within the context of cross-linguistic comparison of ejective stops, but using non-native stimuli (Percival 2023). However, none have specifically focused on the perception of Oromo language stop contrasts, where the manner in which the implosive in particular contrasts with the ejective stop is of interest given that it has been described as glottalic but phonologically voiceless, like ejectives (Lloret 1994).

Methods. 25 first language speakers of Eastern Oromo (12 female and 13 male, aged 17-62 (mean = 42)) participated in a forced-choice identification task where they listened to a series of stimuli and for each clicked on the word that they heard. There were four response types given as choices, as the stimuli were created from a naturally produced minimal quadruplet that differed only in the word-medial coronal stop present: [míit'úú] 'to labour, to deliver baby', [míitúú] 'she who mistreats', [míídúú] 'to comb', and [míɗúú] 'to mistreat'. Each of the four words were divided into three sections, as shown in Figure 1: the preceding vowel and stop closure (which includes the initial [m], and is abbreviated pv+c), the release burst (b), and the following vowel (v). The burst intensity was also manipulated such that there were two versions of each baseline burst: quiet (55 dB) and loud (65 dB). Each of the three pieces were systematically alternated to create all combinations for a total of 128 stimuli (4 baseline preceding vowel and closures x 4 baseline bursts x 2 burst intensities x 4 following vowels) and 3200 listener responses. The responses were statistically analyzed using four linear mixed effects models in R, one model for each response type: response (1 for the response type of interest, 0 for the other response types) ~ burst type + burst intensity + vowel type + preceding syllable & closure type + (1 | Participant).



(a) *Ejective stop: míit'úú* (b) *Implosive stop: míɗúú* (c) *Voiced stop: míídúú* (d) *Voiceless stop: míítúú*

Figure 1: *Oromo baseline stimuli.*

Results. The results are illustrated in Figure 2, where the % response is on the y-axis, the response type chosen by listeners is represented through the different line colours, and the x-axis shows the levels of each manipulated dimension of the stimuli. Across the dimensions, burst type showed the largest effect on listener responses, particularly on the % ejective (red), implosive (yellow), and voiceless pulmonic (blue) responses. For a given response type, the corresponding baseline burst elicited more of that response type than non-corresponding baseline bursts. For example, there were more ejective responses (red line) when the baseline ejective burst was present. As for burst intensity, listeners were significantly more likely to respond ejective (red) or voiceless pulmonic (blue) with high burst intensity, and were significantly more likely to respond implosive (yellow) or voiced (green) with low burst intensity. With vowel type, listeners responded with a given category significantly more when that baseline vowel type was present, except that they surprisingly responded ejective not implosive more when the baseline implosive vowel was present (and significantly more than when the baseline ejective vowel was present). For preceding vowels and closures, listeners had significantly more ejective responses (red) when the baseline ejective preceding vowel and closure was present than when the baseline voiced preceding vowel and closure was present, but did not differ in ejective responses between ejective, voiceless pulmonic, and implosive preceding vowels and closures. They also responded implosive (yellow line) significantly more when the baseline implosive preceding vowel and closure was present compared to any of the other preceding vowel and closure

types. Voiced responses (green) were significantly more likely with voiced preceding vowels and closures, though they were also more likely with implosive preceding vowels and closures as opposed to either voiceless stops and ejectives. Voiceless pulmonic responses (blue) showed the opposite pattern: they were significantly less likely with voiced stop preceding vowels and closures, followed by implosive baseline preceding vowels and closures.

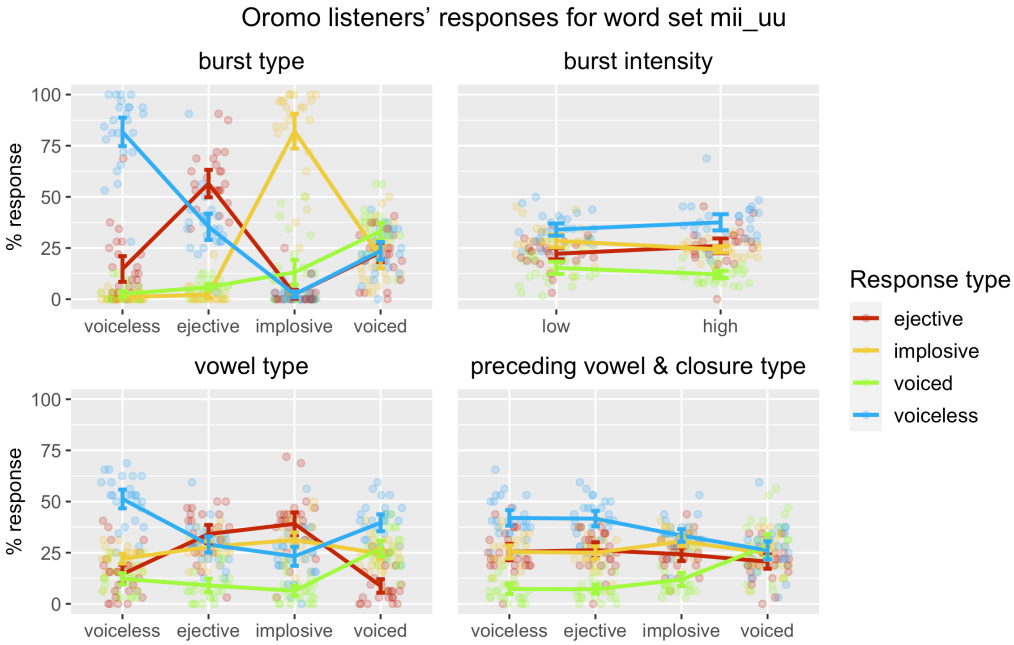


Figure 2: *Results. % response for each response type by dimension of manipulation. % response is given on the y-axis, response type is indicated through line colours, and levels of each dimension are given on the x-axis.*

Discussion. Burst type seems to be a primary cue to the stop laryngeal contrast, particularly for voiceless, implosive, and ejective stops. The acoustic properties of the baseline bursts appear distinct in Figure 1, and so the high usage of burst type is perhaps not unexpected. The ejective baseline burst seems to be shorter and louder compared to the voiceless stop burst, the baseline voiced burst occurs with voicing, and the implosive burst is less distinct compared to the other bursts as they seem to have somewhat affricated releases, likely as a result of being followed by a high back vowel. The results also suggest that the other dimensions may be secondary cues, but whose use is affected by perceptual similarities across certain phonetic features. The first feature is phonetic voicing: listeners associated low intensity bursts with implosive or voiced stops (both with phonetically voiced bursts) but heard high intensity bursts as either ejective or voiceless pulmonic stops (both with phonetically voiceless bursts). This is despite voiced stops having been found to have higher intensity bursts than voiceless and ejective stops in production (Percival 2014). In addition, listener responses did not differ between stimuli with baseline voiceless closures (whether ejective or voiceless pulmonic), but listeners heard more voiced stops for stimuli with baseline voiced and some baseline implosive stop closures (both with voicing in the closure). The second feature is airstream mechanism: listeners broadly responded with either glottalic stop type more when either baseline glottalic stop vowel was present, and in contrast responded with either pulmonic stop more for either baseline pulmonic stop vowel. As for a comparison of the ejective and implosive, the two did differ in cue use. One cue of interest that differs between the two is vowel type, where the baseline implosive vowel sounded more ejective than implosive. This may relate to stimuli acoustics and the relative weighting of vowel type. As seen in Figure 1, the baseline implosive vowel has the most extensive creaky voice of the baseline vowel types. Given that it ended up eliciting more ejective responses, this suggests that creaky voicing may be a slightly more important cue to ejectives than implosives. The implosives and ejectives relying on cues differently suggests that even if the implosive phonologically patterns as voiceless, listeners still perceive voicing differences (as well as potential differences related to airstream).

References

Lloret, M-R. (1994). Implosive Consonants: Their representation and sound change effects. *Belgian Journal of Linguistics*, *9*(1), 59‑72.

Percival, M. (2014). Variation in Ejectives: An Acoustic Study of Stop Contrasts in Eastern Oromo and Dɛlɪnɛ Slavey. MA thesis. University of Toronto.

Percival, M. (2023). Perceptual cues to ejective stops across languages. In R. Skarnitzl & J. Volín (Eds.), Proceedings of the Proceedings of the 20th International Congress of the Phonetic Sciences. Guarant International. 282–286.

Percival, Maida, Alexei Kochetov, and Yoonjung Kang. (2018). An ultrasound study of gemination in coronal stops in Eastern Oromo. In Proceedings of Interspeech 2018.

ARTICULATORY DYNAMICS IN A TONAL LANGUAGE

L Tomui Dangshawa[1]. & Irfana, M[2].

1. Speech Language Pathologist, All India Institute of Speech and Hearing, Manasagangothri, Mysore, India, email ID: kavya11eranhikkal@gmail.com, adithyagopank4@gmail.com & lulusherebi@gmail.com

2. Assistant Professor in Speech Sciences, Department of Speech Language Pathology, All India Institute of Speech and Hearing, Mysore, India, email ID: irfana@aiishmysore.in (Corresponding author)

Abstract

Introduction: The use of contrastive pitch specifications at every level of the phonological hierarchy, in comparison to the latter's use at the segmental level, is the primary distinction between tonal and non-tonal languages. Therefore, the types of segmental and tonal knowledge must be recognized to accurately distinguish a spoken word. Laryngeal mechanism plays major role in tonal change and the way other articulators moves in conjunction with the laryngeal system is indispensable. There have been multiple studies to understand the laryngeal system variation during the production of various tones using acoustical and physiological methods (Shastri & Kumar, 2015; Moisik et al., 2014). However, there is no reported physiological study to understand the articulatory dynamics of any articulator. The Manipuri is a tonal language which has conflicting tonal organization and recent studies highlighted two distinctive tones i.e level and falling tones (Singh, 2019; Devi & Das, 2021). Present study assessed the tongue dynamics during the production of level and falling tones of the same monosyllable utterance that provide an insight regarding the differentiating characteristics of tongue movements if any than the laryngeal features.

Aim and objectives: The study aimed to understand the tongue contours of monosyllabic level and falling tonal counterparts in Manipuri with objective as 1) to obtain tongue contours during the production of monosyllabic level and falling tonal counterparts in Manipuri and 2) to compare the horizontal and vertical tongue dynamics across anterior, mid and posterior tongue regions between level and falling tonal counterparts.

Method

Participants: A total of 10 native speakers of Manipuri language with equal number of males and females selected for the study. All the subjects were more than 18 years of age. All of them were native speakers of Manipuri language (L1) and did not have any history of cognitive, hearing, and speech language impairment. They did not have any other structural abnormalities such as ankyloglossia, macroglossia, cleft lip and palate, had not undergone glossectomy and were not using any dental/oral prosthesis.

Materials: 10 minimal pair of monosyllabic words contrasting in tone were considered for the study (Singh, 2019; Devi & Das, 2021). Each of these words presented in phrases or sentences to produce in various natural tones.

Instrumentation and procedure: The Mindray Ultrasound 6600 and Articulate Assistant Advanced (AAA) software were used to record and analyze the utterances. Before the subject starts the task, the transducer, a long-handled, 6.5MHz micro-convex probe, positioned under the chin. The auditory stimuli recorded by a microphone built into the headset in time with the ultrasound image signal. The acquired images and audio samples were assessed at a 60 frames per second processing rate by the AAA. Individual recordings were done after they are seated comfortably on chair.

Analysis: Tongue contours were analyzed in three points i.e. tongue anterior, mid-dorsum and posterior regions based on the (x-y) coordinates. Anterior (x1, y1) measured where the first spline crossing the tongue contour and mid (x2, y2) the points were plotted where fourth spline crossing the tongue contour and finally posterior (x3,y3) considered seventh tongue spline crossing the tongue contour. Spline kept constant for both level and falling tones to eliminate the variability.

Results: There were significant difference in horizontal tongue dynamics between falling and level tone. Vertical tongue dynamics were not significantly different for the tonal counter parts. Overall, pattern of tongue contour had shifted posteriorly for falling tone by end of the utterance which was not seen for level tone.

Discussion and conclusion: Present study provided insight in to the involvement of articulatory dynamics for the production various tones of Manipuri language.

Lingual articulation of syllabic and non-syllabic /r/

Kateřina Bujoková[1], Tanja Kocjančič[1,2]
[1]Faculty of Arts, Charles University, Prague, Czech Republic
[2]Faculty of Education, University of Ljubljana, Slovenia
katnira.b@gmail.com, tanja.kocjancicantolik@ff.cuni.cz

Introduction. The standard Czech consonant /r/ is a voiced alveolar trill. It is articulated by pressing the edges of the tongue body against the upper molars, or even part of the palate, which leaves the tip of the tongue free to oscillate against the alveolar ridge. Typically, /r/ is realized with 1 oscillation (Machač & Skarnitzl 2009) but older sources state 1 to 3 oscillations (Palková, 1994). The trill is always voiced and is classified as a sonorant, meaning that it has a periodic, tonal structure and easily traceable formants. Furthermore, Czech /r/ can function as a syllabic peak in which case we speak of syllabic or syllable-forming consonants. Traditionally, /r/ becomes syllabic if it is positioned between two other consonants (e.g. *brk "quill"*) or at the end of the word after one or more consonants (*bobr "beaver"*) (Volín, & Skarnitzl 2020). Previous studies conducted in the 1980s did not find any distinctive phonetic contrasts between the syllabic variants. The only reported difference was in quantity, however, this was likely influenced by the neighboring sounds (Bičan 2014). Articulatory examination of syllabic consonants in Berber (based on EPG data, Fougeron & Ridouane 2008) and Slovak (based on EMA data, Pouplier & Beňuš 2011) revealed no difference between the two variants. The authors of the latter study further concluded that syllables with a syllabic consonant behave as consonant clusters, albeit with lesser temporal overlap between consonants.
The current study aims to examine the lingual articulation of syllabic and non-syllabic /r/ in Czech in terms of (1) tongue shape and position and (2) coarticulation. The underlying articulatory mechanisms of both variants can be best studied with ultrasound tongue imaging (UTI). Since the method allows direct observation of the entire tongue contour in the midsagittal plane, it enables observing the differences in the shape and position of the tongue inside the oral cavity (Stone 2005). An UTI study of Catalan (non-syllabic) /r/ has shown tongue blade and predorsum lowering, both necessary for the apical trilling, with tongue body backing in the pharyngeal area (Recasens & Rodríguez 2017). Additionally, the two variants can be described in terms of coarticulation. Earlier UTI studies have shown greater lingual coarticulatory resistance of vowels than consonants (Recasens & Rodríguez 2017; Zharkova & Hewlett 2009). The question remains how this phenomenon is applied to syllabic and non-syllabic varieties of the same sound.

Methods. Six Czech adult native female speakers, aged between 20 to 25 years, participated in the study. All were university students and none reported having speech or hearing difficulties. The word list included 5 disyllabic words with syllabic /r̩/, for example *krky ("necks")*, and 10 disyllabic words with non-syllabic /r/ (5 with /Cr/ sequence in a syllable onset and 5 with /r.C/ sequence over syllable boundary). The C and V environment remained stable which enabled forming of word triplets: e.g. *korky ("corks")*, *krky ("necks"), kroky ("steps"),* or *borky ("small pine groves"), brky ("quills"), broky ("buckshot")*. The word list contained 38 other words used as fillers. Participants made 3 repetitions of randomized target words embedded into a simple phrase *Řekneš __ znovu* ("*Say __ again*"). The audio and ultrasound recordings of midsagittal tongue contour were carried out simultaneously using the Micro system (Articulate Instruments Ltd. 2012) and probe stabilization system (Derrick et al. 2018). To trace the tongue contour, one ultrasound frame per target /r/ was manually selected: the frame right before the opening of the first vibration, i.e. the moment before the tip of the tongue began to fall. Assuming that at the moment of vibration, the tongue, not including the tip, will be the most stable. Mean tongue contours were created from the repetitions of individual speakers. This allowed comparing the lingual articulation of syllabic /r̩/ in different environments and with the non-syllabic variant (by the preceding consonants [kr̩kɪ] x [dr̩bɪ, tr̩sɪ] x [br̩kɪ, pr̩sɪ] and by the following consonants [kr̩kɪ, br̩kɪ] x [tr̩sɪ, pr̩sɪ] x [dr̩bɪ]).

Results. Figure 1 shows the mean midsagittal tongue contours of one speaker. All the contours show a posterior tongue position with lowering in the middle and front part of the tongue. The analysis of the contours of the syllabic /r̩/ showed that the neighboring consonants influence the overall articulation to some extent. Figure 1a includes syllabic /r̩/ in different CrCV words and illustrates these differences. First, it can be seen that in [kr̩kɪ] the tongue is in the highest position along the entire tongue contour, except the front, and the most posterior. The next in terms of height and posteriority is [br̩kɪ], followed by [pr̩sɪ] which has the same height as [dr̩bɪ], however, the latter is produced more anteriorly. The tongue is the lowest in [tr̩sɪ]. The front of the tongue is in the same position for all non-velar environments, while the tongue root is in a similar position for all except [dr̩bɪ]. Figure 1b shows the stability of /r/ in syllabic and non-syllabic variants. Comparing the two triplets differing only in the first consonant suggests that the syllabic variant shows greater resistance to neighboring sounds compared to the non-syllabic variants, particularly in the back of the tongue. First, the tongue contours for the syllabic /r̩/ (preceded by either /k/ or /b/) differ less than the tongue contours of the non-syllabic /r/ forming a consonantal cluster with the same consonants. Second, the bilabial

stop in the CC onset causes more anterior tongue root placement in /r/ than the same stop in the syllabic /ṛ/ context. Third, non-syllabic /r/ shows a uniform effect of the preceding back vowel /o/. Looking at green and blue contours, the sequence in which /r/ is preceded by a vowel and followed by a consonant ([korkɪ] and [borkɪ] in green) has the back of the tongue moved significantly more backward and the middle is slightly elevated compared to the structure in which /r/ is preceded by a consonant and followed by a vowel ([krokɪ] and [brokɪ] in blue).

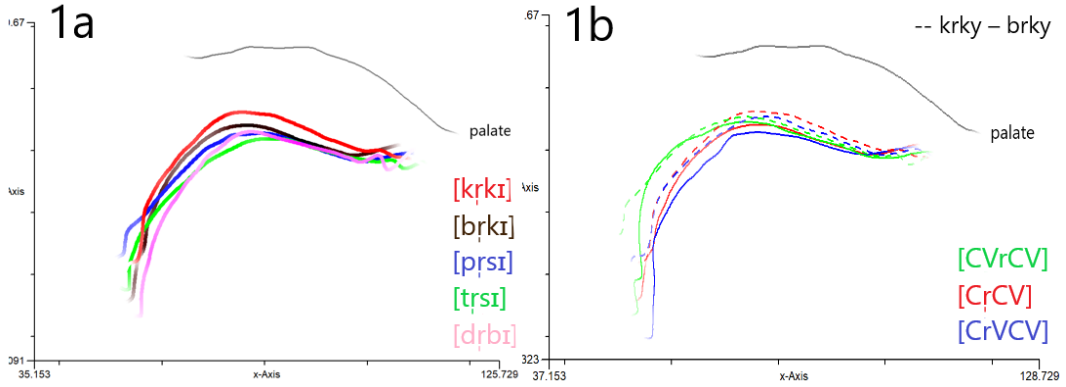


Figure 1: *1a) Mean of midsagittal tongue contours of (1a) a syllabic /ṛ/ in different CrCV words, and (1b) two non-syllabic /r/ and a syllabic /ṛ/. The front of the tongue is on the right side of the figures.*

Discussion. The current study showed that both non-syllabic and syllabic Czech /r/ are articulated with a similar tongue shape as observed in Catalan /r/ (Recasens & Rodríguez 2017) contributing to the articulatory description of this sound. In terms of coarticulatory resistance, the results suggest that syllabic /ṛ/ is more resistant to different phonetic environments than the non-syllabic variety. In the current data set, this is true for cases when /r/ is preceded by a vowel or a consonant, although the exact size of the effect remains to be quantitatively evaluated. The only exception is /r/ preceded and followed by a velar stop where high dorsum in consonant production causes elevated tongue position also for /r/. In terms of articulatory variability of syllabic /ṛ/ in different environments, the tongue contours display the expected patterns, particularly in combination with the alveolar sounds and the following high front vowel. Overall, the coarticulatory resistance is particularly notable in the front part of the tongue. This follows the conclusions by Recasens & Rodríguez (2017) that coarticulation occurs in the parts of the tongue that are not directly involved in the formation of constriction or closure. In /r/ production, the tongue is laterally braced in its middle part (Gick 2017) with the active tongue tip, leaving the root of the tongue rather free and available for coarticulatory changes. Finally, the data shows the presence of a strong carryover coarticulation effect for both /r/ variants. For the syllabic /ṛ/, this is evident in higher back tongue contour in [kṛkɪ] compared to [bṛkɪ], and more anterior tongue root and back in [tṛsɪ] compared to [pṛsɪ]. For the non-syllabic /r/, carryover V-on-r is stronger than anticipatory V-on-r. The study contributes to our understanding of coarticulation, as visible in articulatory data, and the behavior of consonants as syllabic nuclei.

The research was supported by Czech Science Foundation Grant No. 23-05494S.

References

Articulate Instruments Ltd. (2012). Articulate Assistant Advanced user guide: Version 2.14. Articulate Instruments Ltd.
Bičan, A. (2014). Nuclearity of /r/ and /l/ in Czech. In J.Witkós & S. Jaworski (Eds.). New Insights into Slavic Linguistics. Frankfurt am Main: Peter Lang, 21-33.
Derrick, D., Carignan, C., Chen, W. R., Shujau, M., & Best, C. T. (2018). Three-dimensional printable ultrasound transducer stabilization system. *The Journal of the Acoustical Society of America,* 144(5), 392-398.
Fougeron, C., & Ridouane, R. (2008). On the phonetic implementation of syllabic consonants and vowel-less syllables in Tashlhiyt. *Estudios de Fonética Experimental*, 140-175.
Gick, B., Allen, B., Roewer-Després, F., & Stavness, I. (2017). Speaking tongues are actively braced. *Journal of Speech, Language, and Hearing Research,* 60(3), 494-506.
Machač, P., & Skarnitzl, R. (2009). Principles of phonetic segmentation. Prague: Epocha.
Palková, Zdena. (1994) Fonetika a fonologie češtiny. Praha: Karolinum.
Pouplier, M., & Beňuš, Š. (2011). On the phonetic status of syllabic consonants: Evidence from Slovak. *Laboratory phonology*, 2(2), 243-273.
Recasens, D., & Rodríguez, C. (2017). Lingual articulation and coarticulation for Catalan consonants and vowels: An ultrasound study. *Phonetica*, 74(3), 125-156.
Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical linguistics & phonetics*, 19(6-7), 455-501.
Volín, J., & Skarnitzl, R. (2020). Segmentální plán češtiny. Praha: Karolinum.
Zharkova, N., & Hewlett, N. (2009). Measuring lingual coarticulation from midsagittal tongue contours: Description and example calculations using English/t/and/ɑ. *Journal of Phonetics*, 37(2), 248-256.

Positional Effect of Main Stress in Italian: an Articulatory and Acoustic Study

Bowei Shao[1,2], Philipp Buech[2], Anne Hermes[2], Maria Giavazzi[1]

[1]*Département d'études cognitives, École Normale Supérieure, Université Paris Sciences & Lettres*

[2]*Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne Nouvelle, Paris, France*

{bowei.shao;maria.giavazzi}@ens.psl.eu, {philipp.buech;anne.hermes}@sorbonne-nouvelle.fr

Introduction. Lexical stress may be signalled through a large number of acoustic parameters (Gordon and Roettger 2017). In Italian, stress is realized through (1) longer duration, (2) higher intensity, and (3) more peripheral acoustic vowel shape (Albano Leoni et al. 1995; Eriksson et al. 2016; Shao et al. 2023). Duration has been shown to be sensitive to the position where stress occurs in the word: penultimate stressed syllables are longer than antepenultimate stressed syllables (D'Imperio and Rosenthall 1999; van Santen and D'Imperio 1999). Little is known (Avesani et al. 2007) however on the articulatory correlates of this positional effect. This acoustic and articulatory study has three aims: (i) replicating the positional durational difference found in previous studies, (ii) relating the stress-induced differences in intensity to stress-induced differences in lip aperture and jaw movement, and (iii) relating the stress-induced differences in formant structure to stress-induced tongue dorsum positions. The findings will be discussed within the hyperarticulation and the sonority expansion theories of prominence.

Methods. We reanalysed a dataset collected for studying the effect of penultimate stress on post-tonic segments. Target words were trisyllabic nonce words structured /C_1V_1.C_2V_2.C_3V_3/, differing solely by the position of stress on C_1V_1 or C_2V_2. The V_1 and V_3 positions were occupied by /i/. C_1 and C_2 were occupied by /p, t/ and V_2 was /a, e/. C_3 was /k, g, tʃ, dʒ/. (e.g., /ˈpi.ta.ki/, /pi.ˈta.ki/). Nonce words were written in their orthographic forms with lexical stress marked by an accent (e.g., pítachi, pitáchi). They were embedded in a carrier phrase "Pimpa parte da __la mattina presto" and randomized. Fifteen speakers were recorded with EMA (AG501) with synchronised acoustic signals. Only ten speakers were included in the analysis because they had consistent [e] quality when this vowel was stressed and unstressed (10 speakers $\times$ 4 repetitions $\times$ 64 targets = 2560 tokens). Both the tongue dorsum position and the formant structure of the vowels were measured at the mid-50% of their acoustic duration using the window method (WM) (Roessig et al. 2022). Formants were normalized into Bark, and tongue dorsum WM values were z-scored across speakers. To measure intensity, all acoustic files were scaled to have a mean intensity at 55 dB. Then 10 equally distributed points were measured for each vowel as a time-series. Lip aperture was measured as the euclidean distance between the two lips. The statistics were conducted in R using the mgcv and brms packages, reported as GAMM (Generalized Additive Mixed Model) estimate figures, and the mean (with lower and upper boundaries of the 95% credible interval) of the posterior.

Results. Regarding (1) the acoustic duration of stressed vowels, the positional effect is observed in the current data set: penultimately stressed vowels are 40.45 ms ([20.38, 60.68]) longer than antepenultimately stressed vowels, though vowel specific acoustic duration may have played a role. With respect to (2) the relationship between the vowel's intensity and its articulatory aperture, we found a consistent pattern presented in **Figure 1**: across positions, stressed, [i, e, a] all have larger lip aperture (note that GAMMs estimate a significant difference for all three vowels.) In the case of antepenultimate stress, lip aperture is clearly related to the vowel's intensity profile. We can see this relation in **Figure 1** [i], in which the opening-closing pattern of lip aperture in both stressed and unstressed conditions seems to behave in accordance with intensity, in the sense that larger lip aperture could be related to higher intensity. In the case of penultimate stress however, lip aperture is inconsistent with the intensity profile, as can be seen from **Figure 1** [e, a]: the intensity profile does not track the lip aperture one. Furthermore, the dynamic intensity profile shows a important divergence not observed elsewhere: when stress is in the penultimate position, the intensity peak occurs very early in the vowel, followed by a intensity slope (note that we only refer to the overall intensity trajectory shapes). Finally, the analysis of (3) acoustic vowel space and tongue dorsum position reveals a position- and vowel-dependent pattern. For the vowel [i] in antepenultimate position, there is no observable stress related pattern in tongue dorsum position (high-low: $\hat{\beta}$=-0.04 [-0.07, -0.002]; front-back:$\hat{\beta}$=-0.04 [-0.10, 0.03]) or F1 ($\hat{\beta}$=0.03 [-0.14, 0.19]). But the F2 is slightly lowered when it is not stressed ($\hat{\beta}$=-0.20 [-0.32,

-0.08]). In the penultimate position, the tongue dorsum in [a] is lowered ($\hat{\beta}$=-1.06 [-1.22, -0.90]) and fronted ($\hat{\beta}$=-0.64 [-0.84, -0.43]) when the vowel is stressed. As for [e], the tongue dorsum's position shows slight raise along the high-low dimension ($\hat{\beta}$=0.14 [0.02, 0.26]) and a retraction along the front-back dimension ($\hat{\beta}$=0.32 [0.14, 0.50]). Formant structure of penultimate vowels indicates that [a] is lowered ($\hat{\beta}$=1.55 [1.35, 1.75]) and acoustically more posterior ($\hat{\beta}$=-1.37 [-1.71, -1.03]) when it is stressed. The formant structure of [e] is not influenced by stress in F1 ($\hat{\beta}$=0.04 [-0.20, 0.28]) and slightly fronted in F2 ($\hat{\beta}$=0.34 [0.16, 0.52]).

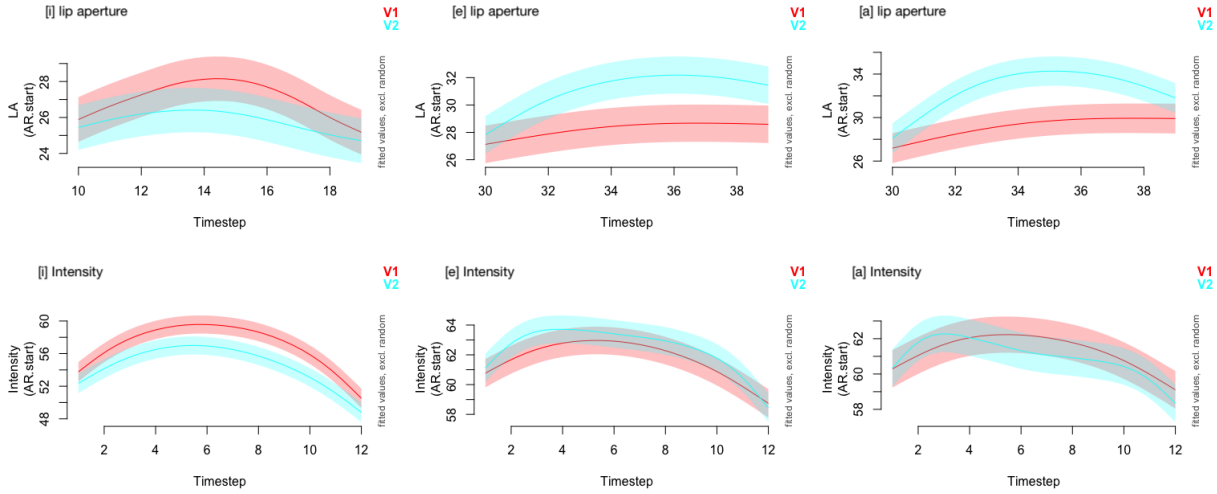


Figure 1: *GAMM estimates of lip aperture (upper) and intensity (lower) during V_1 [i], V_2 [e], and V_2 [a] respectively. Red lines indicate when stress is on V_1 (here [i]), blue lines indicate when stress is on V_2 (here [e, a]).*

Discussion. Our result (1) confirmed earlier acoustic studies on the durational differences between antepenultimate and penultimate stress in Italian. We further characterised the positional effect with respect to articulatory-intensity and articulatory-spectral correlates with result (2). First, the dynamic pattern of intensity, which differed as a function of the stress positions, cannot be directly explained by lip aperture pattern. This observation could be related to a difference in the distribution of perceptual cues to stress in the two positions. Hypothetically, the earlier peak in penultimate stress may be an perceptual cue. Second, with respect to the spectral and articulatory characteristics of stressed vowels, we observed both acoustic and articulatory expansion, attested by the lip aperture for all vowels and further, a lowering of F1 for [a]. Moreover, we also observed that stressed [e, a] are hyperarticulated, as attested by the tongue dorsum position and F2. The current dataset does not have the same vowel in the two stress positions, the ongoing study will allow us to further investigate the systematicity of these dynamic (acoustic and articulatory) differences across a wider range of vowels in comparable positions, and their interactions with surrounding consonants.

References.

Albano Leoni, Federico, Francesco Cutugno, and Renata Savy (1995). "The vowel system of Italian connected speech". In: *Proceedings of the International Congress of Phonetic Sciences*, pp. 396–399.

Avesani, Cinzia, Mario Vayra, and Claudio Zmarich (2007). "On the articulatory bases of prominence in Italian". In: *Proceedings 16th ICPhS 2007*, pp. 981–984.

D'Imperio, Mariapaola and Sam Rosenthall (1999). "Phonetics and phonology of main stress in Italian". In: *Phonology* 16.1, pp. 1–28.

Eriksson, Anders, Pier Marco Bertinetto, Mattias Heldner, Rosalba Nodari, and Giovanna Lenoci (2016). "The acoustics of lexical stress in Italian as a function of stress level and speaking style". In: *Interspeech*, pp. 1059–1063.

Gordon, Matthew and Timo Roettger (2017). "Acoustic correlates of word stress: A cross-linguistic survey". In: *Linguistics Vanguard* 3.1, pp. 1–11.

Roessig, Simon, Bodo Winter, and Doris Mücke (2022). "Tracing the phonetic space of prosodic focus marking". In: *Frontiers in Artificial Intelligence* 5, p. 842546.

Shao, Bowei, Philipp Buech, Anne Hermes, and Maria Giavazzi (2023). "Stress Conditioned Phonological Process: A Case Study of Italian Palatalization". In: *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023)*. Prague, Czech Republic, pp. 2189–2193.

van Santen, Jan and Mariapaola D'Imperio (1999). "Positional effects on stressed vowel duration in Standard Italian". In: *Proceedings of the 14th International Congress of the Phonetic Sciences, San Francisco*, pp. 1757–1760.

Articulatory planning of spoken utterances based on Optimal Control Theory

Benjamin Elie[1], Juraj Šimko[2], Alice Turk[1]

[1]*Linguistics and English Language; School of Philosophy, Psychology and Language Sciences the University of Edinburgh; Edinburgh, Scotland, United Kingdom*

[2]*Department of Digital Humanities; Faculty of Arts; University of Helsinki; Helsinki, Finland*

benjamin.elie@ed.ac.uk, juraj.simko@helsink.fi, a.turk@ed.ac.uk

Introduction.

Speech production requires performing complex overlapping movements of several supralaryngeal articulators (i.e. the lips, the tongue, the velum and the jaw), laryngeal articulators and the lungs. These movements aim at producing acoustic audio signals conveying linguistic and extralinguistic information to the listener. How these movements are coordinated is still subject of debate, as several questions still need to be addressed. They include the nature of representations and processes in planning speech articulations. For instance, models of speech production assume different natures for goals of speech movements. Most dominant models based on Articulatory Phonology (Browman and L. M. Goldstein 1986), such as AP/TD (Saltzman and Munhall 1989), or ETD (Simko and Cummins 2010), assume asymptotic, context-independent articulatory goals, but see Harper, L. Goldstein, and Narayanan (2020) for context-dependent goals drawn from a distribution. In most of these models, contextual differences are assumed to arise from differences in spatial positions at movement endpoints, e.g. via undershoot from shorter activation intervals, gestural blending, or sensitivity to feedback about target achievement. Recently, a model has been proposed that requires full spatial and temporal specification of the realized movement targets (Turk and Shattuck-Hufnagel 2020b; Turk and Shattuck-Hufnagel 2020a). This kind of model uses General Tau theory (Lee 1998) as a dynamic articulatory model, which has been recently shown to fit real articulatory movements better than dominant damped oscillator-based models (Elie, Lee, and Turk 2023). In order to account for contextual variation, such a model requires context-dependent articulatory targets and ways of determining the position and timing of these targets.

This paper presents a model of articulatory planning of spoken utterances. Our approach combines 1) realistic modeling of basic articulatory movements based on General Tau Theory (Lee 1998; Elie, Lee, and Turk 2023), as well as 2) the assumption that speech articulation is an optimal resolution of various cognitive and physical demands (Perrier, Ma, and Payan 2005; Simko and Cummins 2010; Patri, Diard, and Perrier 2015; Parrell and Lammert 2019; Elie, Šimko, and Turk 2023). We will present the theoretical background of the model and preliminary results of simulations that provide an evaluation of a system of high-level control parameters that can be used in phonological and phonetic planning processes. This model is intended to contribute to the debates about the nature of representations and processes in planning speech articulations by proposing an alternative approach for dynamic planning of spoken utterances.

Our model of speech articulatory planning.

In our model, speech production is modeled as an optimization process which satisfies multiple conflicting objectives. The optimization step consists of finding the parameters of the Tau equation of articulatory movement so that the utterance minimizes a cost function which penalizes articulatory effort, low phoneme intelligibility, and long duration utterances. Like (Simko and Cummins 2010; Windmann, Šimko, and Wagner 2015), we propose 3 different objectives: 1) intelligibility, 2) least articulatory effort, and 3) brevity, modeled by the following composite multi-objective function:

$$\mathcal{C}(\theta) = \alpha_{\mathcal{E}}\mathcal{E}(\theta) + \alpha_{\mathcal{P}}\mathcal{P}(\theta) + \alpha_{\mathcal{D}}\mathcal{D}(\theta), \tag{1}$$

where $\mathcal{C}(\theta)$, $\mathcal{E}(\theta)$, $\mathcal{P}(\theta)$, and $\mathcal{D}(\theta)$ are the overall cost, the effort cost, the parsing cost (related to intelligibility), and the duration cost (related to speech brevity), respectively, all functions of the model parameter vector θ. Maximizing intelligibility is predicted to lead to hyperarticulation, whereas minimizing articulatory effort is predicted to lead to hypoarticulation. These conflicting demands can be balanced and modulated by adjusting the weights $\alpha_{\mathcal{E}}$, $\alpha_{\mathcal{P}}$, and $\alpha_{\mathcal{D}}$, assigned to the effort, parsing, and duration costs, respectively.

Our model can account for different prominence levels via the introduction of local weights assigned to specific constituents. These local weights are varied according to the prominence of the constituent with two parameters:

- a positive integer which represents the hierarchical prominence level (the larger the more prominent)
- a real positive number which is used to fine-tune the acoustic effect of prominence (*e.g.*, specific to one language)

Simulations.
We performed simulations of VCV sequences of a specifically customized language composed of a 5 vowels system and 3 voiced stop consonants. These simulations aimed at illustrating the effect of modifying the various parameters of the model and also to verify the ability of our model to reproduce and predict some speech phenomena. Our simulations show that our model is able to predict the centralization and reduction of unstressed vowels, as shown in Figure 1, which shows the vowel spectral density of stressed and unstressed vowels in the $F1 - F2$ vowel space. Similarly, our model predicts the lengthening of stressed vowels.

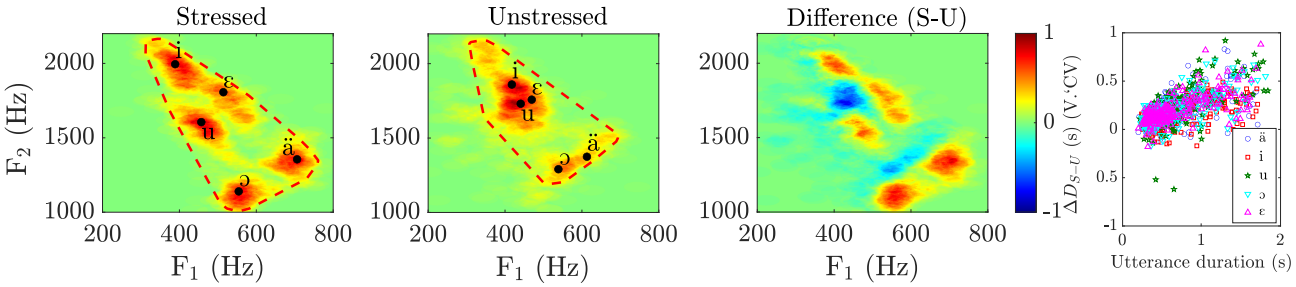


Figure 1: Vowel space density computed for stressed (left plot) and unstressed (middle left plot) vowels. The location of the maximal density of each individual vowel is displayed as a black circle. The convex hull at a density level of 0.25 is displayed as a red dashed line. The middle right plot shows the space density difference between stressed and unstressed vowels. The right plot shows the difference in terms of surface duration between stressed and unstressed vowels (ΔD_{S-U}) as a function of the speech rate, when the stress is on the first syllable.

References.

Browman, Catherine P and Louis M Goldstein (1986). "Towards an articulatory phonology". In: *Phonology* 3, pp. 219–252.

Elie, Benjamin, David N Lee, and Alice Turk (2023). "Modeling trajectories of human speech articulators using general Tau theory". In: *Speech Communication* 151, pp. 24–38.

Elie, Benjamin, Juraj Šimko, and Alice Turk (2023). "Optimal control of speech with context-dependent articulatory targets". In: *Interspeech 2023, Dublin*.

Harper, Sarah, Louis Goldstein, and Shrikanth Narayanan (2020). "Variability in individual constriction contributions to third formant values in American English /ɹ/". In: *The Journal of the Acoustical Society of America* 147.6, pp. 3905–3916.

Lee, David N (1998). "Guiding movement by coupling taus". In: *Ecological psychology* 10.3-4, pp. 221–250.

Parrell, Benjamin and Adam C Lammert (2019). "Bridging dynamical systems and optimal trajectory approaches to speech motor control with dynamic movement primitives". In: *Frontiers in Psychology* 10, p. 2251.

Patri, Jean-François, Julien Diard, and Pascal Perrier (2015). "Optimal speech motor control and token-to-token variability: a Bayesian modeling approach". In: *Biological cybernetics* 109, pp. 611–626.

Perrier, Pascal, Liang Ma, and Yohan Payan (2005). "Modeling the production of VCV sequences via the inversion of a biomechanical model of the tongue". In: *Proceedings of the 9th European Conference on Speech Communication and Technology*. InterSpeech'2005 Editor, ISSN 1018-4074, pp. 1041–1044.

Saltzman, Elliot L and Kevin G Munhall (1989). "A dynamical approach to gestural patterning in speech production". In: *Ecological psychology* 1.4, pp. 333–382.

Simko, Juraj and Fred Cummins (2010). "Embodied task dynamics". In: *Psychological review* 117.4, pp. 1229–1246.

Turk, Alice and Stefanie Shattuck-Hufnagel (2020a). "Speech timing: Implications for theories of phonology, speech production, and speech motor control". In: vol. 5. Oxford University Press, USA. Chap. How do timing mechanisms work?, pp. 238–263.

— (2020b). "Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production". In: *Frontiers in Psychology* 10:2952.

Windmann, Andreas, Juraj Šimko, and Petra Wagner (2015). "Optimization-based modeling of speech timing". In: *Speech Communication* 74, pp. 76–92.

Chewing Efficiency and Oral developmental functions in Children with Oral- and Speech Motor Disorders Compared to Peers

Helena Björelius[1,2], Jonny Trang[3], Fredrik Johansson[1,4], Georgios Tsilingaridis[3,5], Royne Thorman[1,2], Hayo Terband[6]

[1]Karolinska Institutet, Department of Clinical Sciences, Danderyd Hospital, Stockholm, Sweden
[2]Oral Motor Centre (OMC), Division of Speech and Language Pathology, Department of Neurology, Danderyd Hospital, Stockholm, Sweden.
[3]Karolinska Institutet, Department of Dental Medicine, Division of orthodontics and pediatric dentistry, Stockholm, Sweden.
[4]Medical library, Danderyd Hospital, Stockholm, Sweden
[5]Center of Pediatric Oral Health, Stockholm, Sweden
[6]Department of Communication Sciences and Disorders, University of Iowa, Iowa City IA, USA
helena.bjorelius@ki.se, jonny.trang@ki.se, fredrik.johansson.2@ki.se, Georgios.tsilingaridis@ki.se, royne.thorman.1@ki.se, hayo-terband@uiowa.edu

Introduction. Knowledge regarding chewing skills amongst children is limited. Studies demonstrate that the chewing behavior of typically developing children (TD) between 3 to 17 years is similar to adults though chewing efficiency seems to vary especially if there is malocclusion (Almotairy et al. 2021; Alshammari et al. 2022; Mogren et al. 2022). Chewing efficiency in children with oral developmental delay (OD) and motor speech disorders (MSD) under the age of 6 has not been thoroughly investigated (Kaya et al. 2017). Children with MSD are a heterogeneous group that show vulnerabilities such as hypo or hypertonicity of muscles or hypo/hyper-sensitivity of the sensory system, dysfunctions of basic motor patterns like chewing, swallowing, and drooling as well as defects in speech sound production originated from disorders in sensorimotor processing (Björelius & Tükel 2017; Kent 2015; Newmeyer et al. 2009; Nijland et al. 2015). Comorbidity of language disorders (PDL) is also common, altogeher making differential diagnosis between subtypes demanding (Iuzzini-Seigel et al. 2022; Murray et al. 2023). The process of chewing is a complex matter consisting of concomitant neurological, physiological, motor, and sensory activity under strict regulation by central pattern generators allocated in the brain stem (Barlow et al. 2010; van der Bilt et al. 2006). The function can be delayed in children with oral motor challenges and can present itself in refusal of food, piccy eaters, taking in too much food, or swallowing food without chewing (Morris 2000). Failure in chewing efficiency can affect quality of life (Chen & Engelen 2012). The present study investigated the efficiency of mastication in children with MSD compared to children with typical development (TD) using the HueCheck chewing gum test as well as oral developmental challenges and habits.

Methods. The project was part of the regular clinical practice at the Oral Motor Center (OMC), a specialist clinic under the Department of Neurology and Division of Speech Language Pathology at Danderyds Hospital AB, Stockholm Sweden. The study was approved by the Stockholm ethical board, informed consent was obtained from or on behalf of all participants. 285 patients referred to OMC between 4 to 9 years with questions of oral- and speech motor dysfunctions between January 2022 to June 2023 were asked to participate in the study. 240 accepted and 201 fulfilled the assessment (SG group). Of these 201 patients, 111 children agreed to participate in the chewing study (HSG group; 74 boys, 37 girls). Data from a control group (CG group) with 62 (24 boys, 38 girls) between 4 to 9 years of TD children was carried out in a consecutive manner based on their recalls for regular dental examinations at the Dental clinic at Karolinska Institutet, at a private dental clinic, and from colleagues at OMC. Oral sensory and motor challenges were assessed through the anamnesis's information through verbal interviews with the caregivers and from The Swedish translated version (not published) of the Verbal Motor Production Assessment for Children (VMPAC). Chewing efficiency was assessed based on the coloring pattern of the 2-coloured chewing gum after 20 chewing cycles. For the analysis, the gums were placed and flattened to a wafer with a thickness of 1mm using a metal plate with a mild depression of 1 mm x 50 mm x 50 mm. Each of the 111 specimens was photographed from both sides and assessed by a single calibrated operator, experienced with the analysis. The outcome measure SDHue is a color dispersion metric calculated by automated image analysis software. Statistical analyses were conducted through SPSS using Pearson Chi-Square test and multiple linear regression analyses, age separate and gender and age together as covariate. Post hoc tests were performed both at the level of broader groupings of children (MSD; OD) and at the level of specific combinations of diagnoses.

Results. After assessment at OMC, 133 children in the SG group received an OD diagnose and 97 children an MSD diagnose **(**Speech Disorder unspecified with motor origin [ATYP]; Speech Disorder unspecified with motor origin [ATYP with CAS features]; Childhood Apraxia of Speech [CAS]; Dysarthria [DYS]; Oral dyspraxia affecting speech motor control [OA**]**). 118 children were prior assessment at OMC diagnosed with PDL. In the HSG group (*n*=111), 5 children were diagnosed with MSD only, 14 with OD only and 13 with PDL only. 24 had a combination of MSD and OD, 19 with MSD, OD and PDL, 12 with MSD and PDL and 17 with OD and PDL. 7 children are not included in the analyses (5 not receiving a diagnosis and 2 children with ankyloglossia). Descriptive statistics were calculated on the entire study group (SG; *n*=201) for age, gender, and MSD diagnoses and were compared with the children that agreed to do the Hue-Check task (HSG; *n*=111). Descriptive statistics showed satisfactory equivalence regarding age, gender, and diagnoses. Regarding the control group (CG; *n*=62) there was satisfactory equivalence with SG and HSG regarding age, but there were differences in the distribution of gender. Statistical analyses showed that both the children with MSD and the children with a combination of MSD and OD chewed significantly less efficient compared to the controls (all *p*'s<.001). The children diagnosed with OD only did not chew significantly less efficient (*p*=.068). The entire MSD group (*n*=60) as well as the OD group without MSD (*n*=31) chewed significantly less effective than the TD group (*p*=<.001; *p*=.005). Comparison between the entire MSD and OD without MSD groups did not reveal a difference in chewing efficiency (*p*=.527). Regarding oral sensory and motor challenges and habits including *Drooling*, *Babkin reflex*, *Mouth Stimuli*, *Selective eater* and *Stuffs mouth full* children with MSD and OD showed higher frequencies of each symptom compared to the CG group (Chi-Square tests; all *p*'s<.001). Detailed results will be available at the conference.

Discussion. These results indicate that children with MSD show reduced chewing efficiency compared to TD children but not in comparison to children diagnosed with OD and no MSD. The children with MSD also show deviances in oral sensory and motor behaviors and habits including refusal of food and taking in too much food which can depend on delayed chewing (Morris 2000) or possibly be part of a larger symptom complex. Interestingly, the children diagnosed with OD only did not show deviant chewing efficiency compared to TD children which could be influenced by the small sample. As impaired chewing efficiency can affect quality of life (Chen & Engelen 2012), assessment of mastication (and intervention) needs to be considered in clinical practice regarding children with MSD. The present results strengthen earlier studies reporting oral dysfunctions in children with speech sound disorders (Mogren et al. 2022). Speaking and chewing are governed by the same muscular system though there are deviant opinions as whether they are codependent (see e.g., Kent 2015). Further research on coexisting oral motor symptoms and underlying causes in the MSD population is warranted.

References.

Almotairy, N., Kumar, A., & Grigoriadis, A. (2021). "Effect of food hardness on chewing behavior in children". In: *Clinical Oral Investigations* 25, pp. 1203-1216.

Alshammari, A., Almotairy, N., Kumar, A., & Grigoriadis, A. (2022). "Effect of malocclusion on jaw motor function and chewing in children: systematic review". In: *Clinical Oral Investigations* 26.3, pp. 2335-2351.

Barlow, S. M., Radder, J. P. L., Radder, M. E., & Radder, A. K. (2010). "Central pattern generators for orofacial movements and speech". In: *Elsevier Science & Technology* 19, pp. 351-369.

Björelius, H., & Tükel, Ş. (2017). "Comorbidity of motor and sensory functions in childhood motor speech disorders". In: *Advances in Speech-langauge Pathology*: IntechOpen.

Chen, J., & Engelen, L. (2012). *Food oral processing: fundamentals of eating and sensory perception*. Chichester, UK: Wiley-Blackwell.

Iuzzini-Seigel, J., Allison, K. M., & Stoeckel, R. (2022). "A tool for differential diagnosis of childhood apraxia of speech and dysarthria in children: A tutorial". In: *Language, Speech, And Hearing Services In Schools* 53.4, pp. 926-946.

Kaya, M. S., Güçlü, B., Schimmel, M., & Akyüz, S. (2017). "Two-colour chewing gum mixing ability test for evaluating masticatory performance in children with mixed dentition: validity and reliability study". In: *Journal of oral rehabilitation* 44.11, pp. 827-834.

Kent, R. D. (2015). "Nonspeech Oral Movements and Oral Motor Disorders: A Narrative Review". In: *American Journal Of Speech-Language Pathology* 24.4, pp. 763-789.

Mogren, Å., Sand, A., Havner, C., Sjögreen, L., Westerlund, A., Agholme, M. B., & Mcallister, A. (2022). "Children and adolescents with speech sound disorders are more likely to have orofacial dysfunction and malocclusion". In: *Clinical and Experimental Dental Research* 8.5, pp. 1130-1141.

Morris, S. E. (2000). *Pre-feeding skills: a comprehensive resource for mealtime development*. [2nd ed.]. San Antonio TX: Therapy Skill Builders.

Murray, E., Velleman, S., Preston, J. L., Heard, R., Shibu, A., & McCabe, P. (2023). "The Reliability of Expert Diagnosis of Childhood Apraxia of Speech". In: *Journal of Speech, Language, and Hearing Research*, pp. 1-18.

Newmeyer, A. J., Aylward, C., Akers, R., Ishikawa, K., Grether, S., deGrauw, T., Grasha, C., & White, J. (2009). "Results of the Sensory Profile in children with suspected childhood apraxia of speech". In: *Physical & Occupational Therapy in Pediatrics* 29.2, pp. 203-218.

Nijland, L., Terband, H., & Maassen, B. (2015). "Cognitive functions in childhood apraxia of speech". *Journal of Speech, Language, and Hearing Research* 58.3, pp. 550-565.

van der Bilt, A., Engelen, L., Pereira, L. J., van der Glas, H. W., & Abbink, J. H. (2006). "Oral physiology and mastication". In: *Physiology & Behavior* 89.1, pp. 22-27.

Timing of acceleration peaks and acceleration changes

Malin Svensson Lundmark

Lund University, Sweden; Queen Margaret University, UK

malin.svensson_lundmark@ling.lu.se

Introduction. Segment transitions have been shown to consist of *acceleration peaks* of primary active articulators (Svensson Lundmark 2023). In fact, for any speech posture of an active articulator we may find acceleration peaks at the edges of the speech postures, seemingly dividing the postures from the fast intervals of the movements to and from the speech postures (Svensson Lundmark & Erickson 2024).
Mathematically, acceleration is the second derivate to position, and acceleration peaks occur when a mass changes its velocity the most, which it does in connection with changing direction (Eager et al. 2016). In the bottom of **Figure 1** we see the position of an EMA tongue tip sensor of a speaker producing the Swedish word <bilar> (*cars*). As the speaker shapes the tongue tip constriction in /l/, the tongue tip moves fast (a *velocity peak*) and then slows down rapidly (a *deceleration peak*). The tongue tip stays in position while forming the speech posture of /l/, and then moves rapidly away again (an *acceleration peak*, followed by a *velocity peak*).

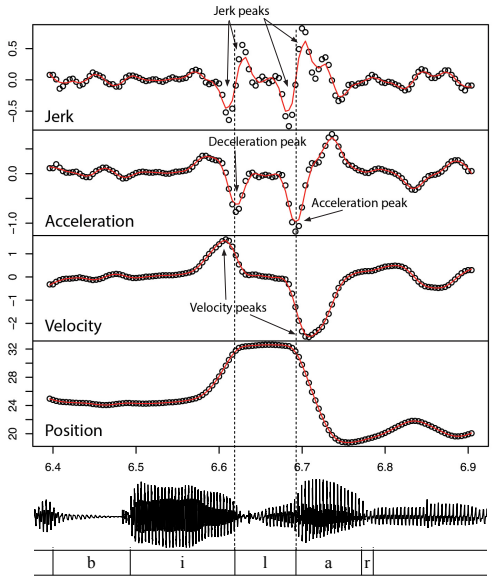


Figure 1. *The vertical position of tongue tip, with velocity, acceleration and jerk, while producing /l/ in the word <bilar>. Red solid lines are the smoothed EMA signals. Speech signal and segments in the bottom rows.*

On the top row in **Figure 1** you find jerk (third derivate to position). A jerk peak occurs when acceleration changes the most (Eager et al. 2016), and these appear on either side of the de/acceleration peaks as acceleration changes both before and after its maximal value (**Figure 1**). The acceleration peaks and the acceleration changes (jerk peaks) coincide with the edges of an articulatory posture and in extension with the segment boundaries (vertical dotted lines in **Figure 1**). Recent studies show that this relationship between de/acceleration peak and acoustic segment boundary is robust and holds across e.g. syllable strength, prominence levels, tonal context, and manner and place of articulation (Svensson Lundmark 2022, 2023; Svensson Lundmark & Frid 2023; Svensson Lundmark & Erickson 2024). This study reports on some of these findings on different articulators (lips, tongue, mandible), and discusses these results in light of acceleration changes (jerk peaks) in supraglottal articulation and the DASA approach (Descriptive Approach to Segmental Articulations; Svensson Lundmark & Erickson 2024).

Method. The subsets of data on which results are reported are from a corpus with 18 South Swedish speakers recorded with an EMA system (Carstens AG501, 250 Hz) at the Lund University Humanities Laboratory. Speakers read from a prompter leading questions and target sentences with disyllabic target words, each set displayed eight times in random order. EMA position data was collected from a number of articulators (see detailed information in Svensson Lundmark 2023). Here is reported on sensors on lips (lip aperture), tongue tip, tongue dorsum, and lower incisors (lower jaw). Post-processing of signals was done in Carstens software, and in R, where specifically calculation and articulatory analyses

were performed. The acceleration was derived by computing the second-order differences of the position data using a time window of 0.02 seconds. The acceleration signal has been filtered and smoothed using a low-pass filter, the R function *loess*. Acoustic segmentation was done by the author in Praat (an IAA was also performed, see details in Svensson Lundmark 2023). Collection of de/acceleration landmarks of word-initial CVC sequences (consisting of open vowel /a/, and /m/ and /n/) were done semi-automatically in R (landmarks were visually inspected and adjusted when justified). Timing of acceleration and deceleration peaks are calculated by measuring the time lag to the expected corresponding acoustic segment boundary. As a statistical tool to evaluate the time lags, linear mixed effects models (LMM) were used and run in R (details in Svensson Lundmark 2023; Svensson Lundmark & Erickson 2024).

Table 1. *Results on average time lags in ms (stdv) of de/acceleration peaks to acoustic segment boundaries of CVC sequences. A negative time lag indicates that the de/acceleration peak precedes the segment boundary.*

	Time lags					
Articulators	C1 onset/ Deceleration	C1 offset/ Acceleration	V1 onset/ Deceleration	V1 offset/ Acceleration	C2 onset/ Deceleration	C2 offset/ Acceleration
Lip aperture	11 (5)	4 (10)	-	-	10 (4)	-2 (8)
Tongue tip	12 (10)	5 (8)	-	-	10 (6)	-2 (7)
Tongue dorsum	-	-	35 (20)	-25 (20)	-	-
Lower jaw	20 (15)	-15 (15)	40 (15)	-45 (20)	20 (12)	-17 (15)

Results. Table 1 shows an overview of the findings on timing of deceleration peaks to acoustic segment onset, and of acceleration peaks to segment offset (as previously reported in Svensson Lundmark 2023; Svensson Lundmark & Erickson 2024). For the primary articulator of a constriction (as lips in /m/, and tongue tip in /n/) we see short time lags at both onset and offset, and in both C1 and C2 position (**Table 1**). The deceleration peak seems to follow the boundary slightly at C1 and C2 onsets. Furthermore, the acceleration landmarks of tongue dorsum are not aligned to V1 onset and V1 offset; we find long and varied time lags indicating a much shorter speech posture than vowel segment (**Table 1**). Note that the acceleration peak of the primary articulator at C1 offset, and the deceleration peak at C2, determine the acoustic vowel segment duration. The lower jaw displays overall long and varied time lags; timing of the de/acceleration peaks tell us that the jaw speech postures are shorter than the postures of the other articulators (**Table 1**).

Discussion. The results on acceleration peak timing of primary consonantal articulators (lips and tongue tip), tongue dorsum, and lower jaw, paint a rather structured and robust picture. The speech postures of the consonantal articulators, delimited by deceleration and acceleration peaks, shape the segment durations, while the speech postures of the lower jaw appear to be much shorter (that jaw opening begins before lip opening has previously been reported by Fujimura 1961). Similarly, tongue dorsum acceleration peaks shape a short speech posture, but vowel articulation needs further investigation; its complex dynamic behavior may not be captured sufficiently by one EMA sensor.
Why deceleration peaks lag behind the onset segment boundaries may be because of a coordination with the acceleration change (=jerk peak) rather than the acceleration peak (**Figure 1**). This in turn may indicate that the nature of acoustic segment transitions differs between onset and offset, i.e. deceleration and acceleration of articulatory movements.
Using acceleration peaks we may build a descriptive model on speech articulation (the DASA approach; see Svensson Lundmark & Erickson 2024) which predicts that speech postures and fast intervals of all articulators are timed with one another. Consequently, this leads to a structure where acceleration peaks and acceleration changes (jerk peaks) in the orofacial region are timely connected. Furthermore, such a structure also predicts that speech postures consist of two distinctive and jerky movements, one decelerating before changing direction towards the next position, and one accelerating after. Ongoing research aims to map the timing and magnitude of these acceleration peaks and acceleration changes in the orofacial region.

References

Eager, D., Pendrill, A.-M., and Reistad, N. (2016). "Beyond velocity and acceleration: Jerk, snap and higher derivatives," Eur. J. Phys. 37(6), 065008.

Fujimura, O. (1961). Bilabial stop and nasal consonants: a motion picture study and its acoustical implications. J Speech Hear Res. 4, 233–47. doi:10.1044/jshr.0403.233.

Svensson Lundmark, M. (2022). Evidence of segmental articulations: Acceleration determines vowel segment duration in Swedish Word Accents. Proceedings of 1st International Conference of Tone and Intonation (TAI 2021), SDU, Sønderborg.

Svensson Lundmark, M. (2023). Rapid movements at segment boundaries. J. Acoust. Soc. Am. 153 (3), 1452–1467. https://doi.org/10.1121/10.0017362

Svensson Lundmark, M., & Erickson, D. (2024) Segmental and syllabic articulations: a descriptive approach. J. Speech. Lang. Hear. Res. https://doi.org/10.1044/2024_JSLHR-23-00092

Svensson Lundmark, M. & Frid, J. (2023). Segmental articulations across prosodic levels. In: O. Niebuhr & M. Svensson Lundmark (eds), Proceedings of the 13th Nordic Prosody Conference: Applied and Multimodal Prosody Research (pp. 255-261), Sonderborg, Denmark. Warsaw: Sciendo/de Gruyter, https://doi.org/10.2478/9788366675728-023