13th International Seminar of Speech Production

Proceedings

13 – 17 may 2024 Autrans FR

https://issp24.sciencesconf.org

Proceedings of the 13th International Seminar of Speech Production

(ISSP 2024)

Foreword

Since 2011, ISSP's submission process is based on a 2-page abstract. The abstracts that were accepted by the Scientific Committee of ISSP2024 are gathered in the Book of Abstracts. In addition, there is the possibility for each author to submit a 4-page paper. This is not mandatory. This year, 73 papers were submitted. They are gathered in these Proceedings. Importantly, these papers were not reviewed in details. They are conceived as an extension of the accepted abstracts.

A selection of these papers will be proposed for a Special Issue in Speech Communication to be launched. Once selected these papers will undergo the usual review process of this journal. The selection of the papers will be done by an international committee. To date, the project has not been yet finalized. Things will be more concrete after the conference.

Cécile Fougeron & Pascal Perrier

Chairs of ISSP2024

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

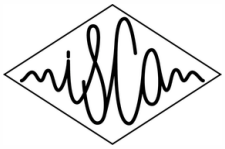

Sociétés et Humanités
Université Paris Cité


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

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.

Table of Contents

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

Abstract

This study examines the constriction geometry of a typologically rare 6-way place contrast in Malayalam nasals. This is done using static MRI data obtained from two speakers. The measures of tongue constriction angle (higher for more posterior places) and tongue constriction length (higher for laminals and dorsals) were found to provide a relatively good characterization of the contrast. Altogether, all pairs of consonants were statistically distinguished by a combination of these variables except for the dental vs. alveolar contrast (and two other pairs for one of the speakers). The nasal consonants were realized as apico-laminal dental /n̪/, apical alveolar /n/, laminal alveolar or alveolopalatal /ɲ/, subapical palatal (retroflex) /ɳ/, fronted velar /ŋʲ/ and (plain) velar /ŋ/. This is largely consistent with previous phonetic descriptions of the sounds and earlier palatographic data available for coronals. Finally, the results for dental and retroflex nasals are compared to similar consonants in Kannada (another Dravidian language), pointing to potential language-particular differences in the realization of the contrast.

Keywords*: speech production, place contrasts, nasals, MRI, Malayalam*

1. Introduction

Malayalam (Dravidian) exhibits a typologically unusual 6-way place of articulation contrast in lingual nasal consonants (Kumari 1972; Asher & Kumari 1997; Namboodiripad & Garellek 2017). As illustrated in Table 1, the contrast involves a series of nasals that are traditionally described as dental, alveolar, retroflex, (alveolo)palatal, and two velars – palatalized and plain. How exactly this complex set of contrasts is distinguished by speakers, however, is unclear.

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

Place	Word	Gloss
dental	pan̪:i	*pig*
alveolar	kanː i	*a month*
retroflex	kaɳ:i	*link*
(alveolo)palatal	kaɲ:i	*gruel*
palatalized velar	matːaŋʲːa	*pumpkin*
(plain) velar	taŋ:i	*held fast*

The only previous articulatory investigation of a subset of these consonants, the coronals /n̪, n, ɳ, ɲ/, was conducted by Dart & Nihalani (1999). Based on static palatograms and linguograms obtained from nine speakers, the authors concluded that the four consonants could be classified into three rather than four places of articulation: denti-alveolar /n̪/, alveolar /n/ and /ɲ/, and postalveolar for /ɳ/. Of note is their finding of a more anterior than expected production of /ɲ/, traditionally described as (alveolo)palatal (e.g., Kumari 1972; Asher & Kumari 1997). The authors also observed that the consonants were differentiated by four constriction shapes: apical for /n/, apico-laminal for /n̪/, apico-sublaminal for /ɳ/, and laminal for /ɲ/. In other words, the four-way contrast in Malayalam coronals was distinguished by a combination of the constriction location (/n̪/ > /n/, /ɲ/ > /ɳ/) and the spatial extent of the constriction (being minimal for apicals and maximal for (sub)laminals).

In this study we examine the constriction geometry of Malayalam nasals using static MRI data from two speakers. In doing this, we are expanding on the tongue tip constriction angle method proposed by Proctor, Bundgaard-Nielsen, Best, Goldstein, Kroos, & Harvey (2010), designed to model a 4-way coronal contrast in Wubuy, an Australian Aboriginal language. In that study, the contrast between laminal dental, apical alveolar, apical retroflex, and laminal alveolopalatal was defined as a series of spatial tongue tip/body targets as angles along a polar grid line of the vocal tract, from the upper teeth to the pharynx, spanning a range of 140°.

We recently adapted this approach to capture the dental-retroflex contrast in Kannada (Dravidian), using static MRI recorded from two speakers (Kochetov, Savariaux, Lamalle, Noûs, & Badin 2024). The results showed that – among stops, nasals, and laterals – the contrast was clearly distinguished by smaller angles (more posterior constrictions) for retroflexes compared to dentals. For example, the dental nasal /n̪/ in the /a_a/ context was produced by the two speakers with a constriction at 156° or 157°, while the angle for the retroflex /ɳ/ was 130° or 126°, respectively. In addition, there were constriction length differences, with higher values for laminals and subapicals compared to apicals (with the types varying by both place and manner).

Unlike Kannada, which has only two coronal and one velar nasal, the set of relevant consonants in Malayalam is considerably larger. It thus remains to be seen whether the tongue constriction angle method is applicable to the complex set of contrasts in Malayalam nasals.

2. Methods

2.1. Speakers, procedure, and materials

Single slice mid-sagittal MRI static images were recorded for two native speakers of Malayalam (SV, female; BB, male; 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/ in five symmetric V_V contexts: /a_a/, /i_i/, /u_u/, /e_e/, and /o_o/ (e.g., [aŋa], [iŋi], [uŋu], [eŋe], [oŋo]). They did it three times in a row, first producing the VCV word twice naturally and then repeating it again and sustaining the articulation of the consonant for about 6.5 seconds. The MRI recordings were taken during the sustained articulation phase. This resulted in a total of 60 images of target consonants (6 consonants x 5 vowel contexts x

1 repetition x 2 speakers). The data were collected as part of a larger corpus of Malayalam sounds.

2.2. Segmentation and tongue constriction geometric characteristics

Semi-automatic segmentation of the main speech articulators from the MRI images was performed according to Labrunie, Badin, Voit, Joseph, Frahm, Lamalle, Vilain, & Boë (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 of this procedure 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.

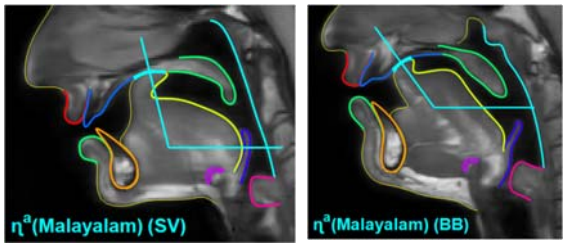


Figure 1: *Articulator contours superimposed on a midsagittal image of /ɳ/ in /aɳa/ by speakers SV and BB with the angle representing the constriction location measure.*

2.3. Statistical analysis

Although our dataset is relatively small, we chose to provide an exploratory statistical analysis of the data. This was done Linear Mixed Effects Regression (LMER) models with tongue constriction parameters TCL and TClength, separately for each speaker. Place (with 6 levels) was a fixed effect, while Vowel (with 5 levels) was a random effect (with random intercepts). The analysis was implemented with the *lme4* package (Bates *et al.* 2015) using *R* (Team, 2014). In each case, likelihood ratio tests were used to compare a full model to a nested model excluding the factor of interest, employing the *Anova()* function of *lmerTest* package (Kuznetsova *et al.* 2017). Pairwise comparisons and post-hoc tests (with a Bonferroni correction for multiple comparisons) were performed using the *phia* package (De Rosario-Martinez 2015).

3. Results

3.1. Overview

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 (159.8°) to the velar place (46.5°); the constriction length is relatively small for the anterior consonants produced with the tongue tip, blade, or the underside (e.g., 1.11 cm for /n/), and is much larger for the posterior consonants produced with the tongue front/body or dorsum (e.g., 3.47 cm for /ŋʲ/). The realization of the first three consonants by the speaker can be described as an apico-laminal dental (or denti-alveolar), apical alveolar, and a subapical palatal retroflex, respectively. The last consonant in the figure is a fairly posterior velar or uvular; /ɲ/ and /ŋʲ/ are fairly similar, differing in the relative frontness of the constriction and the involvement of the tongue dorsum. They can be classified as laminal alveolopalatal and lamino-dorsal palatal. Similar realizations in the /o_o/ context were exhibited by speaker SV, with the exception of /ɲ/, which was produced at a more anterior location (thus being a laminal alveolar). In addition, this speaker's retroflex showed a more retracted constriction (at the dome of the palate; cf. Figure 1).

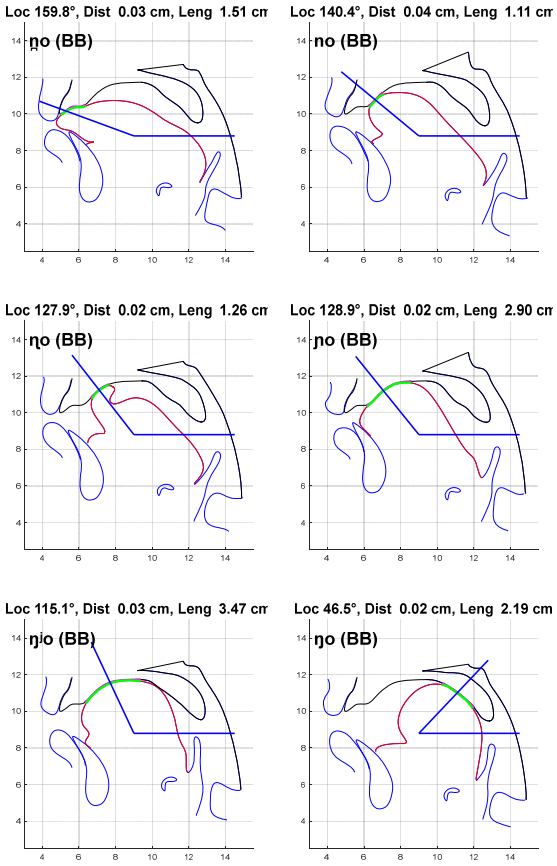


Figure 2: *Constriction location plots for nasal consonants in the /o/ context by speaker BB. Loc: TCL; Dist: Constriction diameter (not used in the study); Leng: TClength.*

3.2. LMER results

Results of LMER models and posthoc tests performed separately by variable and speaker are summarized in Table 2 and are further illustrated in Figure 3. We can see that for speaker SV, TCL angle distinguished dental and alveolar nasals (higher angle) from the retroflex and the two velars; the latter two also differed from each other. The alveolopalatal nasal, however, did not significantly differ in TCL from the other nasals, apart from the plain velar. This can be attributed to the fairly extensive linguopalatal contact for /ɲ/, spanning the alveolar and postalveolar regions. Overall, eight out of 15

pairwise comparisons were significant. For speaker BB, significant TCL differences involved 12 out of 15 pairwise comparisons: all consonants were differentiated from each other apart from the pairs /n̪/-/n/, /n/-/ɲ/, and /ɳ/-/ɲ/. The results for TClength were similar for both speakers: values were significantly higher for the alveolopalatal and two velars compared to the dental, alveolar, and retroflex (with the exception of /ɳ/-/ŋ/ for SV). In other words, the length results reflected differences between laminals and dorsals on the one hand and apicals (and apico-laminals) and sub-apicals on the other.

Table 2: *Results of LMER model comparisons for Tongue Tip Constriction Location (TTCL) and Constriction Length (TTlength) and pairwise posthoc comparisons by speaker (DF = 5).*

	Variable	F	Pr(>F)	Posthoc differences (p<.05)
SV	TCL	18.24	<.001	n̪, n > ɳ, ɲʲ, ŋ; ɲʲ > ŋ; ɲ > ŋ
	TClength	12.77	<.001	ɲ, ŋʲ, ŋ > n̪, n; ɲ, ŋʲ > ɳ
BB	TCL	41.25	<.001	n̪, n > ɳ, ɲʲ, ŋ; n̪ > ɲ; ɳ > ɲʲ, ŋ; ɲ > ŋʲ, ŋ
	TClength	40.38	<.001	ɲ, ŋʲ, ŋ > n̪, n, ɳ

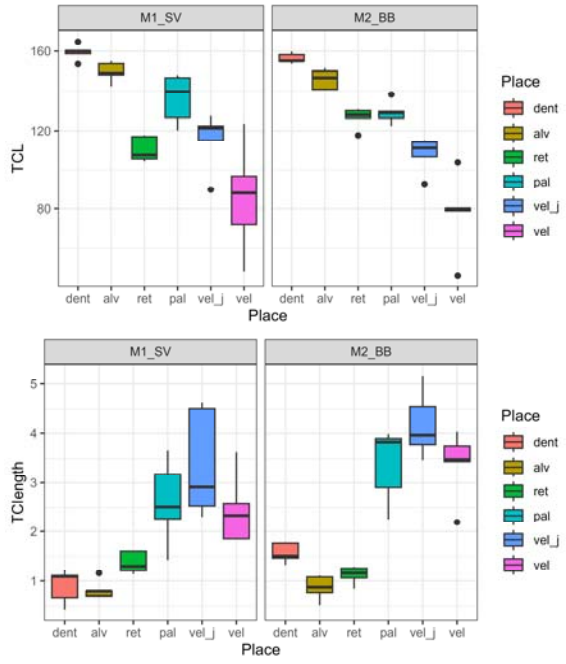


Figure 3: *Boxplots for Tongue Constriction Location (top) and Tongue Constriction Length (bottom) by place (dental, alveolar, retroflex, palatal, palatalized velar, and velar) for both speakers.*

Taken together, all nasals were distinguished by the combination of two measures for speaker BB, with the exception of the dental-alveolar pair (14 out of 15 pairwise comparisons). For speaker SV, the measures distinguished all consonants except for the dental-alveolar, alveolopalatal-palatalized velar, and retroflex-velar pairs (/n̪/-/n/, /ɲ/-/ɲʲ/, and /ɳ/-/ŋ/; 12 out of 15 comparisons). Thus the only contrast that was not distinguished by TCL and TClength measures across the speakers was the dental-alveolar contrast. It should be noted that TCL values for both speakers were on average higher for /n̪/ than /n/ (Figure 3; see also Figure 2). Similarly, TCL values for /ɲ/ produced by SV were higher than /ŋʲ/, as well as higher for /ɳ/ than /ŋ/. The lack of significance in these cases can be due to the relatively small number of analyzed tokens, the overall proximity of dental and alveolar locations, and – for SV – overall greater contextual variability of the data (as evident in larger confidence intervals for most consonants; Figure 3). We assume that the between-speaker differences observed here reflect different individual strategies; however, we cannot exclude the possibility of gender-specific differences.

4. Discussion and conclusion

The goal of this study was to explore the highly complex set of place contrasts in Malayalam by adapting the tongue constriction angle method previously used for coronals in Wubuy (Proctor *et al.* 2010) and Kannada (Kochetov *et al.* 2024). This was done by analyzing static MRI data obtained from two speakers of the language. The results showed that measures of TCL angle and TClength can potentially distinguish all Malayalam nasal contrasts except for the two anterior coronals – the dental /n̪/ and the alveolar /n/. As mentioned above, the lack of significant differences here (which is in contrast to Proctor *et al.* 2010's findings for Wubuy) is likely due to the relative proximity of TCL values for the two consonants, compounded by the small number of items. Two other consonant pairs produced by SV were not clearly differentiated by the measures – /ɲ/-/ŋʲ/ and /ɳ/-/ŋ/.

This suggests that – for a fuller analysis of Malayalam place contrasts – the constriction geometry measures need to be complemented by articulatory modeling parameters, such as tongue tip fronting, tongue body, and tongue dorsum PCA components. These measures, together with TCL and TClength, were used in our analysis of the dental-retroflex contrast in Kannada (Kochetov *et al.* 2024; following the method proposed by Badin, Bailly, Revéret, Baciu, Segebarth, & Savariaux 2002). Tongue tip fronting in particular differentiated dentals from retroflexes, and may therefore be useful in capturing the greater tip protrusion for the Malayalam /n̪/ (relative to both /n/ and /ɳ/. Similarly, the tongue body and dorsum components should contribute to differentiating between retroflex, alveolopalatal, and palatalized and plain velar places. We are currently exploring this approach to Malayalam nasals.

As was found for Wubuy (Proctor *et al.* 2010), coronals in Malayalam are characterized by a combination of place (constriction location) and constriction shape (apical, laminal, subapical). Our two speakers showed overall similar realizations of the consonants, with the exception of /ɲ/, which was produced as laminal alveolar by SV (in contrast to laminal alveolopalatal by BB). Recall that this was also the description of this consonant by Dart & Nihalani (1999) based on palatographic data. The more posterior realization of this consonant by speaker BB, however, points to a larger variability in the production of this consonant than previously observed. Further, our results for the other coronals are consistent with Dart & Nihalani's (1999) conclusions about the articulation of /n̪/ (apico-laminal dental), /n/ (apical alveolar), and /ɳ/ (subapical postalveolar or palatal). Both speakers produced the retroflex /ɳ/ with a considerable curling of the tongue tip (especially for SV) and a large sublingual cavity. This is similar to what has been observed for the Malayalam lateral in the MRI study by Narayanan, Byrd, & Kaun (1999).

As TCL and TClength measurements used here were also employed for Kannada dental and retroflex consonants in Kochetov *et al.* (2024; as noted above), it is worth comparing those measurements (for nasals) to the current results. It can be seen in Figure 4 (top) that the dental /n̪/ was produced as fairly front in both languages (two speakers were pooled in each group). The closure for the retroflex /ɳ/, on the other hand, was considerably more posterior in Malayalam than in Kannada. Further, while the Kannada speakers showed some TClength differences (more linguopalatal contact for the retroflex), the two consonants were relatively similar for the Malayalam speakers. Interestingly, the more retracted articulation of Malayalam retroflexes can be plausibly attributed to its more crowded coronal inventory. This is difficult to ascertain, however, given the small sample sizes of both studies. Future work should explore the potential differences in the production of similar contrasts across Dravidian languages.

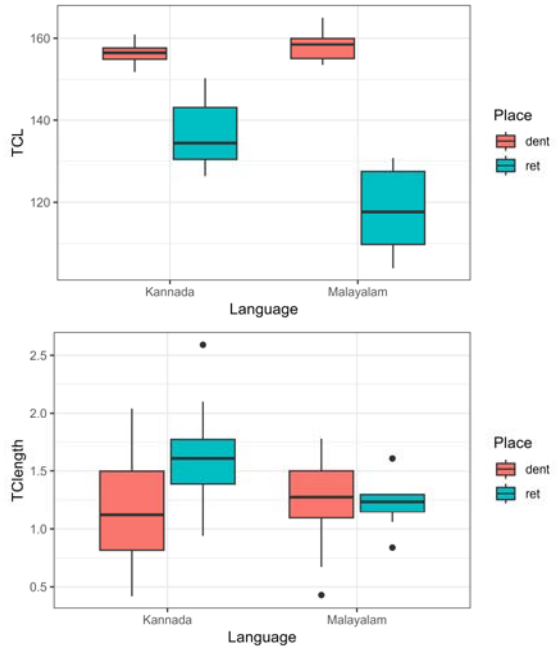


Figure 4: *Boxplots for TCL (top) and TClength (bottom) for dental and retroflex nasals in Kannada (from Kochetov et al. 2024) and Malayalam (current study) (two speakers pooled for each language).*

It is important to note that while the constriction geometry method was originally developed for coronal contrasts, it is clearly applicable to more posterior lingual sounds, such as velars. Interestingly, the palatalized velar in our Malayalam data was more similar in its constriction location to coronals (and especially to /ɲ/) than to its plain velar counterpart. This shows that the consonant is better characterized as a fronted velar ([ŋ̟]) or a dorsal palatal, rather than a velar with a secondary palatal articulation. It should however be kept in mind that our understanding of these consonants is based on static images of sustained articulations. As palatalized consonants typically involve asynchronous coordination of primary and secondary gestures (Kochetov 2006; Shaw, Oh, Durvasula, & Kochetov 2021), dynamic MRI is needed to further investigate this question.

To conclude, our investigation of articulatory properties of a complex set of Malayalam nasals has revealed that the contrasts can be relatively successfully characterized by a combination of two variables – the Tongue Constriction Location angle and Tongue Constriction Length. Further work, however, is needed to provide a fuller characterization of this complex contrast, as well as to investigate possible speaker-/gender-specific strategies and language-particular differences in the realization of similar contrasts.

5. Acknowledgements

The authors would like to thank the two speakers and Christophe Savariaux and Laurent Lamalle for providing extensive support for the data collection and analysis. The data were collected at the IRMaGe MRI facility in Grenoble. The research was partly funded by an Insight Grant from Social Sciences and Humanities Research Council of Canada (#435-2015-2013) to Alexei Kochetov. The MRI facilities centre IRMaGe in Grenoble, France, was partly funded by the grant 'Infrastructure d'avenir en Biologie Santé - ANR-11-INBS-0006' from the French *Agence Nationale de la Recherche*.

6. References

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., & Savariaux, C. (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics 30(*3), 533–553.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using {lme4}. *Journal of Statistical Software, 67*(1), 1–48.

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

De Rosario-Martinez, H. 2015. Package 'phia'. Retrieved from https://github.com/heliosdrm/phia.

Kochetov, A. (2006). Syllable position effects and gestural organization: Articulatory evidence from Russia. In L. Goldstein, D. Whalen, & C. Best (Eds.), *Laboratory Phonology 8* (pp. 565–588). Berlin, New York: De Gruyter Mouton.

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.

Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1-26.

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.

Narayanan, Shrikanth S., Dani Byrd & Abigail Kaun. 1999. Geometry, kinematics, and acoustics of Tamil liquid consonants. *The Journal of the Acoustical Society of America 106*(4), 1993–2007.

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.

Shaw, J. A., Oh, S., Durvasula, K., & Kochetov, A. (2021). Articulatory coordination distinguishes complex segments from segment sequences. *Phonology, 38*(3), 437–477.

Team, R.D.C. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, http://www.R-project.org.

Speech onset kinematics predict sentence level variability in adults who stutter

Torrey M Loucks[1], Daniel Aalto[2,3]

[1]Jacksonville University, Department of Communication Sciences and Disorders
[2]University of Alberta, Department of Communication Sciences and Disorders
[3]Misericordia Community Hospital, Institute for Reconstructive Sciences in Medicine
tloucks@ju.edu, aalto@ualberta.ca

Abstract

The onset of speech involves a state change that could be susceptible to disfluency and higher overall kinematic variability in adults who stutter (AWS). However, the relationship between the susceptibility and kinematic variability has not been explained. In this study, we evaluated the kinematic profile of the initial syllable of a multisyllabic utterance in relation to the variability of the whole phrase. 27 AWS displayed reductions in kinematic amplitude and peak opening velocity and higher initial variability of the lower lip opening gesture compared to 26 adults who speak fluently (AWF). A regression model indicated that initial syllable dynamics of the lower lip predicted higher phrase variability for the AWS group but not the AWF. Atypical speech onset kinematics could propagate into higher overall kinematic variability across an utterance in stuttering.

Keywords: stuttering, spatiotemporal index, speech onset, kinematics

1. 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; Buhr & Zebrowski, 2009; Saltuklaroglu, Kalinowski, Robbins, Crawcour, & Bowers, 2009). Even if speech is initiated fluently however, kinematic variability across a multisyllabic utterance is also higher in AWS and children who stutter (CWS) compared to typically fluent speakers (Loucks et al., 2022; Smith et al., 1995; MacPherson, & Smith, 2013). We hypothesize there is a link between the susceptibility to stutter at speech onset and the high kinematic variability of multiword utterances. Our first step towards testing this hypothesis is assessing whether there is a link between speech onset kinematics and phrase level variability.

Previous reports indicate the onset of fluent utterances in AWS can differ relative to fluent speakers. van Lieshout et al. (1993, 1996) reported delays in lower lip (LL) surface EMG along with increased amplitude at speech onset in AWS. Guitar et al. (1988) reported atypical sequencing of LL EMG sequencing at speech onset in AWS. Additionally, there is substantial evidence that AWS have delayed speech onset times (Bloodstein & Bernstein Ratner, 2008). Altered speech onset dynamics have not been formally related to susceptibility to stuttering disfluencies at speech onset; however, if kinematic variability of the initial syllable is higher in AWS, it could suggest instability of the initial sensorimotor state at speech onset.

We posit that speech onset motor aberrations could dispose a stuttering speaker to higher variability across the whole phrase. Numerous reports have now established that kinematic variability of multisyllabic real word phrases - quantified by the spatiotemporal index (STI) - is greater in both AWS and CWS for multisyllable real word phrases and complex nonwords (Smith, Sadagopan, Walsh, & Weber-Fox, 2010; Smith, Goffman, Sasisekaran, & Weber-Fox, 2012; MacPherson & Smith, 2013).

In this report, we tested our hypothesis by determining whether speech onset kinematics of a multisyllabic utterance are related to whole phrase kinematic variability. We compared LL opening kinematic point measures and variability of the first syllable with the STI of a multiword utterance in AWS and AWF. Then we developed a regression model to test whether the initial syllable dynamic profile can predict phrase level variability. The broader theoretical motivation for this research is to explore whether speakers who stutter experience a sensitivity to initial conditions in which speech can bifurcate into fluency or disfluency.

Prediction 1: Speech onset kinematics of a multiword utterance will differ in AWS compared to AWF.

Prediction 2: The dynamics of the first syllable will predict the spatiotemporal index (STI) of the multiword utterance.

2. Methods

Speech kinematic data were collected in 27 AWS (18-33 years; 20 males) and 26 AWF (19-36 years, 18 males), who are all native speakers of English. All of the AWS were diagnosed with stuttering as children and reported previous treatment for fluency as a child/teen or young adult. None of the AWS were receiving therapy at the time of the study. Stuttering severity was Mild for 9 participants, Moderate for 13 participants and Severe for 5 participants. The percentage of syllables stuttered (%SS) during speaking ranged from 2% SS – 24% SS with a mean of 9.9% SS. The %SS during reading ranged from 1% SS to 23% SS with a mean of 11.8% SS. The experiment was approved by the Ethics Review Board at the University of Alberta (Pro00075834).

The participants were seated comfortably in front of a computer monitor in a quiet room. Each participant produced 15 repetitions of different multisyllabic real word phrases and nonword phrases in randomized order. The initial syllable characteristics and whole phrase variability of one fluently produced multiword utterance - 'Buy Bobby a Puppy' or BBAP - are reported here (Note: Only fluent tokens of the phrase were included). The acoustic signal was acquired with a head-worn microphone at a 2" mouth-to-mic distance (44,000 s/sec). Kinematic recordings of head, jaw upper lip, lower lip and chin motion were acquired with the OptoTrak system (100 s/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 fluent tokens for each participant (inferior-superior dimension) to delineate the phrase (see Figure 1). To determine the variability of the initial syllable ‘buy’, the same onset point was used while the offset point was the peak closing amplitude into the initial /b/ in ‘bobby’. The LL kinematic vectors of the whole phrase and first syllable were then normalized separately in the spatial and temporal domains to 1000 points. The standard deviation was then obtained at 50 points from the matrix of LL vectors to generate separate estimates of the STI for the phrase and initial syllable following the methods of Smith et al. (1995). In a separate analysis, a set of kinematic points of LL in the inferior-superior dimension for the first syllable ‘buy’ were labelled. The onset of /b/ in ‘buy’ was again used as the initial point while the peak opening amplitude of the following vowel served as the second point to identify the following variables using automated algorithms (Figure 1): A) LL opening duration (sec), B) peak LL opening displacement (mm), and, C) peak LL opening velocity (mm/sec) obtained from the first derivative of displacement. Figure 2 shows the overlapped, normalized vectors of LL motion for one AWF participant and one AWS participant over 10 productions of the first syllable ‘buy’.

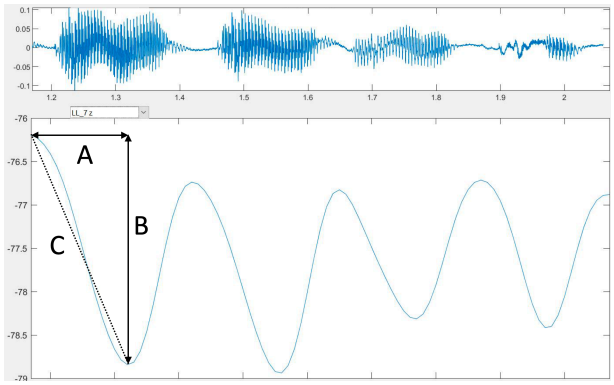


Figure 1. *A representative example of the utterance "Buy Bobby a puppy". The top panel depicts the acoustic signal as a function of time and the bottom panel the Inferior-Superior dimension of the lower lip in head normalized coordinates. The arrow A shows lower lip opening duration, arrow B depicts lower lip opening amplitude, and the slope of the dashed line C represents the peak velocity of lower lip opening.*

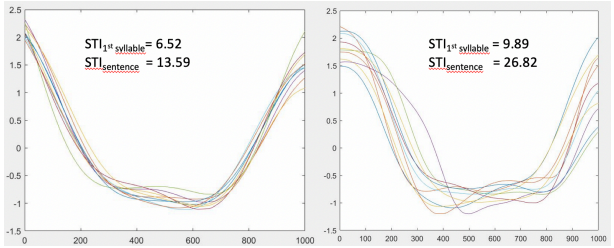


Figure 2. *Time and amplitude normalized lower lip movement trajectories (Inferior-Superior Dimension) for a AWF participant (left) and AWS participant (right). These trajectories are used for calculating the STI of the initial syllable. The STI of the full phrase is also shown.*

3. 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 LL 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 (amplitude, velocity, duration, and variability) were fitted in a regression model to predict whole phrase variability. 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 criterion) with initial syllable STI and/or peak velocity. Residual plots do not suggest violations of model assumptions or undue impact of single observations.

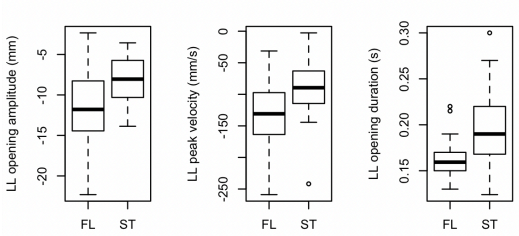


Figure 3. *Average kinematic point measures and standard deviation of LL opening amplitude (mm), LL peak velocity (mm/s), and LL opening duration (s) are shown for the AWS and AWF groups.*

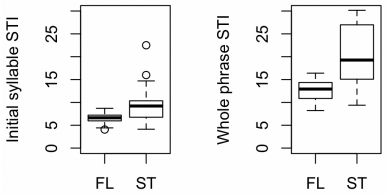


Figure 4. *Average STI and standard deviation of the initial syllable and whole phrase are shown for the AWS and AWF groups.*

Table 1. *Correlations between phrase level STI and first syllable characteristics and clinical data for AWF and AWS. %SS is percentage syllables stuttered for speaking and reading.*

Spearman correlation with sentence STI	AWF		AWS	
	rho	p	rho	p
Lip opening amplitude	0.16	0.43	0.49	0.0097*
Lip opening velocity	0.15	0.48	0.35	0.07
Lip opening duration	0.19	0.35	0.43	0.026*
First syllable STI	0.26	0.21	0.62	0.0008*
%SS in speaking	-	-	0.4	0.039*
%SS in reading	-	-	0.31	0.12

4. Discussion and Conclusions

In this study, the speech onset pattern of AWS involved slower and smaller amplitude LL movements that diverged into more variable kinematic trajectories of the initial syllable and across the whole phrase compared to AWF. These findings support both of our predictions suggesting that motor aberrations at speech onset can predispose AWS to kinematic instability of the whole utterance.

The current findings of altered kinematic point measures at speech onset have not been reported frequently for multiword utterances. The slower LL opening time in the AWS can be related to Van Lieshout et al. (1996) who reported delays in LL EMG at speech onset. The reduced speech amplitude and lower velocity of LL opening suggests a more restricted kinematic approach to speech initiation in AWS relative to AWF. Our measurement of initial syllable variability with the STI extends these findings in showing that aberrations in LL displacement and velocity of the AWS can lead to higher overall speech initiation variability (Figure 4) and potential instability. A renewed focus on speech initiation patterns will be important in future kinematic studies of stuttering.

AWS may have adopted a more cautious and slower speech initiation pattern to avoid disfluencies or due to previous therapy involving speech onset strategies. Faster and larger cautious movements could dispose the person who stutters to higher movement variability at onset, but this would indicate a difference in their approach to speech, rather than a deficit. Alternately, the higher onset variability and altered onset kinematics prompt consideration of whether anomalies in speech planning induce a vulnerability to breakdown at speech onset. Several prominent and recent neurological investigations point to atypical brain activity that precedes the onset of both fluent speech and stuttering disfluencies in AWS (for example, Sengupta et al. 2017; Mersov et al. 2016). Speech planning limitations have long been considered to be involved in the higher susceptibility to stuttering disfluencies at speech onset shown by persons who stutter. These findings of initial syllable instability in AWS potentially point to another aspect of how alterations in prespeech activity could impact speech initiation in stuttering. Deviations at speech onset and variability across the phrase are consistent with an underlying sensorimotor integration deficit in stuttering.

Another novel finding is that the altered speech onset patterns predicted higher phrase level STI in the AWS but not AWF. This finding is challenging to explain in numerical and conceptual terms, because there is no precedent for why kinematic point measures, such as peak displacement, are related to overall phrase variability. One previous study focused on the relationship between kinematic point variables and the STI. Smith & Kleinow (2000) related LL kinematic measures of opening displacement, opening velocity and opening duration to the STI of the BBAP phrase. They reported some AWS tended to have lower amplitude and lower velocity of LL motion compared to AWS, but there were no statistical group differences and these kinematic variables did not predict STI. Our study differed by focusing on the initial syllable, whereas Smith & Klienow (2000) assessed the kinematic opening characteristics of later syllables ('bob' & 'pup'). The current study also had substantially more participants potentially conferring statistical power. The Smith and Kleinow report (2000) did explore other interesting aspects of the potential relationship by varying speech rate but these manipulations did not elicit significant correlations between kinematic point measures and the STI. Despite not identifying a relationship, this study is important for focusing on dynamic interactions between individual kinematic gestures and overall utterance variability.

The specific regression relationship is that larger scale and more rapid movements lead to higher STI in the AWS. This could be consistent with how slow controlled speech onsets became part of therapy approaches. Yet, it's also puzzling because the AWF had larger amplitude and rapid LL motions, which should be destabilizing. While the AWF could generally have more robust speech planning/programming mechanisms that are maintained across the duration of the message, it does not explain the finding in AWS. The AWF group also showed substantial inter-speaker variation for each measure so the non-significant regression did not arise from a restricted range effect.

The STI was developed in the context of dynamical systems theory that viewed stuttering as product of multiple interacting domains and nonlinear interactions that can shift a stuttering speaker from stability to instability (Smith & Weber, 2017). This perspective offers compelling implications for how higher STI values in stuttering speakers at the syllable and phrase level influence susceptibility to speech breakdown. One element of dynamic systems theory is that the behavior of a complex system (e.g., speech) can be sensitive to initial conditions. One set of initial conditions can dispose the system to stable behavior, however, slight changes in initial conditions can also lead to chaotic behavior. As most stuttering occurs at speech onset, research in the context of dynamical systems should focus on identifying and quantifying initial conditions that lead to instability. These considerations were part of the original motivation for this study. We are not certain whether LL opening kinematics represent initial conditions within this perspective but as quantitative measures they can be related to dynamic variables, such as the STI. Computational models of speech kinematics could offer insight into the complex interplay between onset kinematics and sentence level coordination of movement (Simko & Cummins 2011). In particular, the higher STI values in AWS could be due to alternating between different stable production patterns and the onset kinematic pattern (e.g., reduced LL opening amplitude) could be an attempt to lock into a specific production mode.

The significant relationship between speech opening kinematics and the higher STI of AWS suggests correlated structure emerges in their speech production that renders upcoming speech dependent on previous speech (at least within a single utterance). Within this line of thinking, atypical LL opening dynamics could shift the emerging utterance away from stable movement patterns. Yet, it could indicate the speech of typically fluent individuals is not sensitive to initial conditions or has a broader range of stability that allows initial conditions to vary without disposing the system to chaotic trajectories. While the dynamic systems approach has been applied to speech for many years, the consideration of initial conditions in stuttering has not been formally tested. We consider this study as one step towards explaining the 'why' of how kinematic aberrations can occur at speech onset and propagate as higher variability across an utterance. The work on recurrent quantification analysis (RQA) by Jackson and colleagues is related to our discussion because RQA reveals trajectories towards stable or chaotic attractors (Jackson, Tiede, Beal, & Whalen, 2016).

This study provides novel findings that kinematic point measures of speech initiation are related to global variability in

stuttering. Previous therapy could be influencing this relationship if stuttering speakers have adopted strategies for starting speech with smaller and slower movements. An alternate consideration is that aberrant speech onset kinematics and higher overall kinematic variability point to a sensorimotor integration deficit. From dynamical systems theory, the speech of persons who stutter could be particularly sensitive to initial conditions that can bifurcate to fluency or disfluency. Future research should be broadened to investigate: 1) syllable variability within longer phrases, 2) additional utterances, 3) more repetitions, 4) LL EMG, 5) other articulation patterns (e.g., pre/post therapy), 6) rate variations, and 7) intonation variations.

5. Acknowledgements

We are grateful to the participants and the research volunteers who assisted with data collection.

6. References

Bloodstein, O., & Bernstein Ratner, N. (2008). *A handbook on stuttering*. Clifton Park, NY: Delmar

Buhr, A., & Zebrowski, P. (2009). Sentence position and syntactic complexity of stuttering in early childhood: a longitudinal study. *Journal of fluency disorders*, *34*(3), 155–172. https://doi.org/10.1016/j.jfludis.2009.08.001

Guitar, B., Guitar, C., Neilson, P., O'Dwyer, N., & Andrews, G. (1988). Onset sequencing of selected lip muscles in stutterers and nonstutterers. *Journal of speech and hearing research*, *31*(1), 28–35. https://doi.org/10.1044/jshr.3101.28

Jackson, E. S., Tiede, M., Beal, D., & Whalen, D. H. (2016). The Impact of Social-Cognitive Stress on Speech Variability, Determinism, and Stability in Adults Who Do and Do Not Stutter. *Journal of speech, language, and hearing research: JSLHR*, *59*(6), 1295–1314. https://doi.org/10.1044/2016_JSLHR-S-16-0145

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

MacPherson, M. K., & Smith, A. (2013). Influences of sentence length and syntactic complexity on the speech motor control of children who stutter. *Journal of speech, language, and hearing research : JSLHR*, *56*(1), 89–102. https://doi.org/10.1044/1092-4388(2012/11-0184)

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

Saltuklaroglu, T., Kalinowski, J., Robbins, M., Crawcour, S., & Bowers, A. (2009). Comparisons of stuttering frequency during and after speech initiation in unaltered feedback, altered auditory feedback and choral speech conditions. *International journal of language & communication disorders*, *44*(6), 1000–1017. https://doi.org/10.1080/13682820802546951

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

Simko, J., & Cummins, F. (2011). Sequencing and optimization within an embodied task dynamic model. *Cognitive Science*, *35*(3), 527-562. doi: 10.1111/j.1551-6709.2010.01159.x

Smith, A., Goffman, L., Sasisekaran, J., & Weber-Fox, C. (2012). Language and motor abilities of preschool children who stutter: evidence from behavioral and kinematic indices of nonword repetition performance. *Journal of fluency disorders*, *37*(4), 344–358. https://doi.org/10.1016/j.jfludis.2012.06.001

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

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.

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), 1–18. doi: 10.1016/j.jfludis.2009.12.001

Smith, A., & Weber, C. (2017). How Stuttering Develops: The Multifactorial Dynamic Pathways Theory. *Journal of speech, language, and hearing research : JSLHR*, *60*(9), 2483–2505. https://doi.org/10.1044/2017_JSLHR-S-16-0343

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

van Lieshout, P. H., Peters, H. F., Starkweather, C. W., & Hulstijn, W. (1993). Physiological differences between stutterers and nonstutterers in perceptually fluent speech: EMG amplitude and duration. *Journal of speech and hearing research*, *36*(1), 55–63. https://doi.org/10.1044/jshr.3601.55

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

Abstract

The aim of the present study was to test a novel, non-invasive method to measure the middle ear muscle reflex (MEMR) during self-vocalization compared to when presented with external speech, to allow the quantification of its contribution to the reduced response in auditory cortex found in speaking-induced suppression (SIS). MEMR responses were measured utilizing the principle of change in impedance and thus reflectance characteristics of the eardrum, quantified by amplitudes of continuously presented low-level click sounds. Data from 15 healthy young adult speakers show 1) that MEMR was activated prior to the onset of speech, but not prior to the playback of the recorded utterances and 2) that MEMR is stronger for self-generated sound. As MEMR reduces low-frequency excitation in the cochlea, SIS may need to be corrected for MEMR.

Keywords: *speech production, speech synthesis*

1. 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, error correction and speech-motor learning.

Largely ignored in the field of speech production but well-studied in audiology are two major efferent feedback pathways to the auditory periphery: the middle ear muscle reflex (MEMR) reflex and the medial olivocochlear reflex (MOCR; see **Figure 1**). These lower-order CD mechanisms, particularly MEMR, are

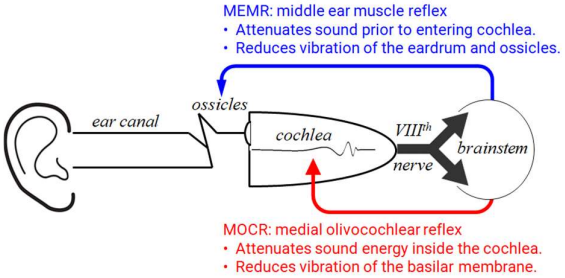


Figure 1: *Two lower-order mechanisms of corollary discharge: the middle ear muscle reflex (MEMR) reflex and the medial olivocochlear reflex (MOCR).*

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, which in turn reduces the input to the cochlea at these frequencies. In quiet environments such as experimental lab conditions, reduced input changes 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, electromyography (EMG) data have 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 for SIS measurement. The current study features a novel, non-invasive method to measure MEMR that allows isolation and quantification of the MEMR component of SIS.

2. Methods

2.1 Participants

Fifteen 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.

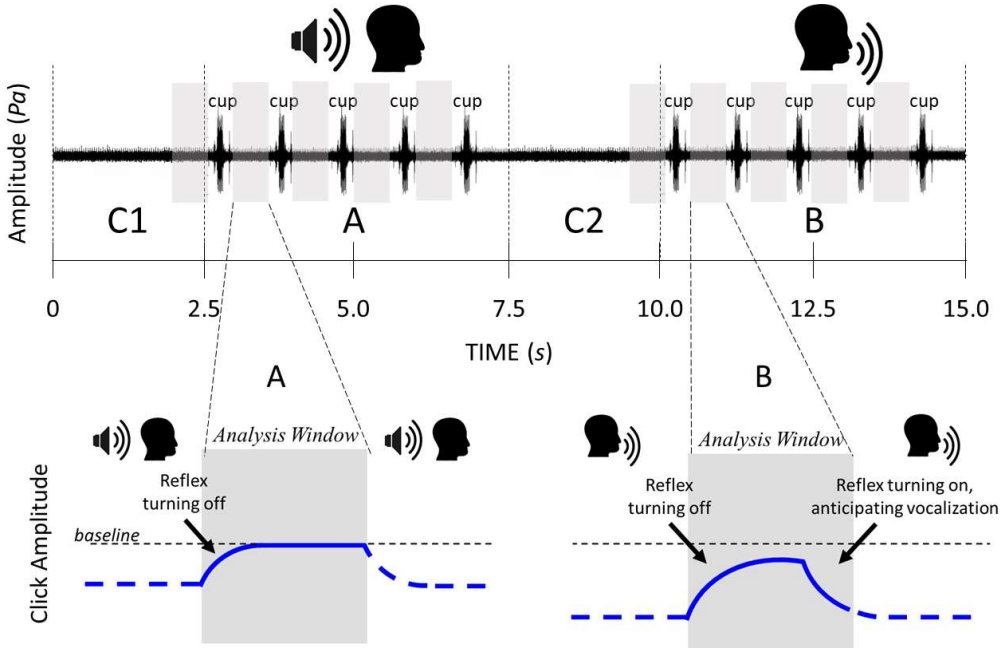


Figure 2: *Experimental paradigm and predicted results. In response to external sound (A), the reflexes will turn on after onset, with a delay of ~100ms. The reflexes will turn off after the sound stops with the same delay. In response to self-produced sound (B), the reflexes will be activated more strongly and prior to onset.*

2.2 Procedures

Participants were seated in a sound-treated booth; stimuli were presented, and ear canal pressure measured binaurally using a 2-channel probe-microphone system (ER10X, Etymotic Research). The experiment consisted of 110 trials, each consisting of four conditions: *Listen*, in which subjects listened to the recording of their own voice playing back the word "cup" five times, a*Speak*, in which subjects were visually cued to produce the word "cup" five times, with a 2.5s inter-stimulus interval, and two *Baseline* conditions. During both *Listen* and *Speak* trials, a train of low-level clicks were played continuously (**Figure 2**). Sound pressure levels of stimuli were equivalent. Baseline conditions, containing clicks only, were interleaved with the *Listen* and *Speak* conditions. The stimulus played in the Listen condition was an ear canal recording of the participant speaking, adjusted so that ear canal sound pressures during Listen and Speak conditions were approximately the same.

2.3 Data processing and analysis

The time-courses and magnitudes of MEMR responses were quantified by measuring the changing amplitudes of the click sounds reflected by the eardrum during the inter-stimulus intervals, compared to the baseline conditions.

3. Results

Figure 3 presents changes in recorded waveform magnitude averaged over five 100 ms time windows during the inter-

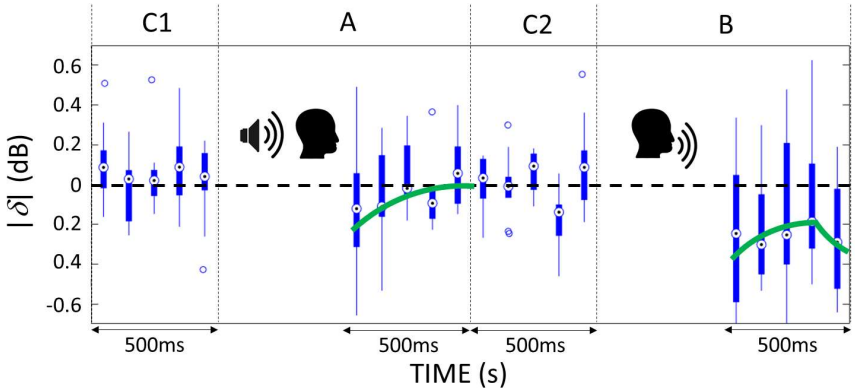


***Figure 3**: Changes in recorded waveform magnitude (|δ|) as a function of time within each condition. At each time point, box and whisker plots show the median changes of the group of participants (N=10; see Fig. 5 to observe all individual data points). Open circles with black dots in the center show the group median values. Thicker blue bars show the second and third quartiles. Open blue circles indicated individual participant median values greater or less than 1.5 times the interquartile range. Exponential trend curves of MEMR activation (green lines) show the changing median activation across the 500 ms analysis windows in the conditions Listen (A), Speak (B) compared to baselines (C1 & C2).*

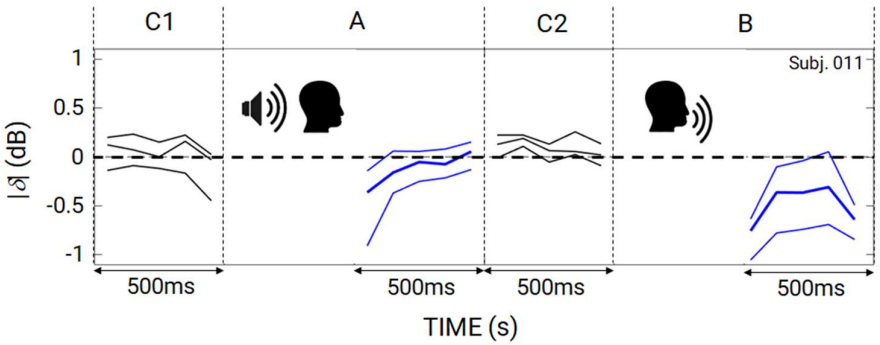


***Figure 4**: Example of MEMR activation curves for an individual participant (subj011) as a function of time within each condition. Median changes in recorded waveform magnitude (|δ|) are shown by the thicker center lines, with interquartile range being shown by slightly thinner lines above and below.*

stimulus intervals in *Listen* and *Speak* conditions compared to baselines. An example of MEMR activation curves for an individual subject is presented in **Figure 4**.

Because the data had a less than optimal signal-to-noise ratio, median values were considered when fitting and reporting the data. The filled and open circles in **Figure 5** show the median values of the 10 participants at five time points in each of the four conditions. In each condition, the data for the five-point time series were fit with a straight line (N=50 data points). The Matlab Curve Fitting toolbox (R2023b) was used, along with the least absolute residuals (LAR) robust fitting option. The LAR method minimizes the absolute difference of the residuals, rather than the squared differences, so that extreme values have a lesser influence on the fit. For the two control conditions (C1 and C2), the 95% confidence intervals for the slopes and the intercepts contained zero. For the two test conditions (A and B), the 95% confidence intervals for the slopes contained zero, but the intercepts did not. Taken together, the results suggest the test conditions produced decreases in ear canal magnitude compared to baseline. Further, the intercept for condition B (-0.376 dB) was outside of the 95% confidence interval for condition A (-0.307 dB to -0.029 dB), highlighting that the *Speak* condition (condition B) produced a larger effect of the MEMR than the *Listen* condition (condition A).

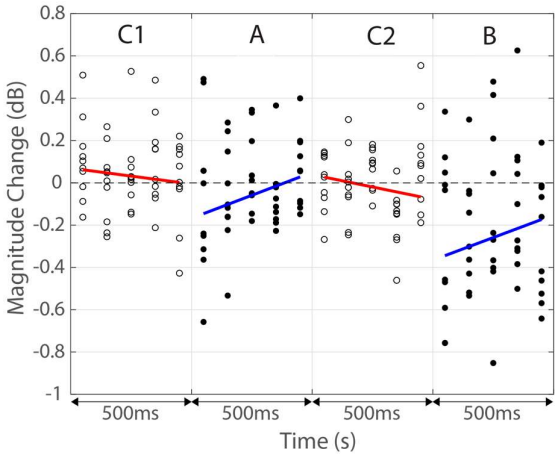


***Figure 5**: Changes in recorded waveform magnitude (|δ|) as a function of time within each condition Each circle shows data from a participant. Open circles show control conditions, and filled circles show test conditions. The red and blue lines show best fits to the data.*

4. Discussion and conclusion

Auditory self-monitoring plays an important role in the acquisition and maintenance of intelligible speech. An important mechanism in auditory self-monitoring is SIS, the phenomenon of a reduced response in auditory cortex to self-produced compared to externally-produced speech, putatively triggered by the efference copy from motor regions (Knille et al. 2019; Ylinen et al. 2015). SIS has been found to be deviant in a variety of prevalent disorders such as stuttering (e.g., Beal et al., 2011; Toyomura et al., 2020) and Parkinson's disease (Mollaei et al., 2019; Huang, et al., 2016) and is thought to play an important role in the online adaptation of vocalizations to environmental conditions and speech-motor learning. However, speech studies have focused on auditory cortical activation while alternative peripheral mechanisms of corollary discharge have been largely ignored.

The present study investigated the MEMR, a peripheral mechanism of that might be of particular relevance in this context. The MEMR involves the contraction of the middle ear muscles, which stiffens the ossicular chain and reduces the low-frequency input to the cochlea. Intracellular EMG of action potentials in the stapedius muscle has shown that the MEMR occurs during (and in anticipation of) vocalization and is stronger during self-vocalization than when presented with external speech (Borg & Zakrisson 1975). Since this peripheral mechanism is altered during self-vocalization, it would have bottom-up effects on cortical activity and thus may alter auditory self-monitoring and play a role in speech production, acquisition and relevant disabilities.

The current study explored a novel, non-invasive method that estimates the magnitude of the MEMR by measuring the sound pressure of acoustic clicks that are reflected by the tympanic membrane in-between speech stimuli. The results show that MEMR magnitude can be assessed non-invasively, both during listening and during self-vocalization. Furthermore, the results were consistent with the EMG findings of Borg & Zakrisson (1975) that the MEMR 1) is activated prior to the onset of speech, but not prior to the playback of the recorded utterances, and 2) is stronger for self-generated sound. These findings potentially have specific implications for research into SIS. As MEMR reduces the excitation amplitude in the cochlea and subsequently the response in auditory cortex, measurements of SIS may need to correct for the effect of the MEMR. Experiments involving simultaneous measurements of MEMR and cortical activation for self-vocalizations are being prepared.

5. Acknowledgements

The authors would like to express their gratitude to all participants thank for their time and effort in this study. Additionally, we would like to thank two anonymous reviewers of our conference proposal for their constructive comments.

6. References

Beal, D. S., Quraan, M. A., Cheyne, D. O., Taylor, M. J., Gracco, V. L., & Luc, F. (2011). Speech-induced suppression of evoked auditory fields in children who stutter. *Neuroimage*, *54*(4), 2994-3003.

Borg, E., & Zakrisson, J. E. (1975). The activity of the stapedius muscle in man during vocalization. *Acta oto-laryngologica*, *79*(3-6), 325-333.

Crapse, T. B., & Sommer, M. A. (2008). Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, *9*(8), 587-600.

Greenlee, J. D. W., Jackson, A.W., Chen, F., Larson, C.R., Oya, H., Kawasaki, H., et al. (2011) Human Auditory Cortical Activation during Self-Vocalization. *PLoS ONE 6*(3): e14744.

Herrmann, B., Augereau, T., & Johnsrude, I. S. (2020). Neural responses and perceptual sensitivity to sound depend on sound-level statistics. *Scientific Reports*, *10*(1), 9571.

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

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: an MEG study. *Journal of cognitive neuroscience*, *14*(8), 1125-1138.

Huang, X., Chen, X., Yan, N., Jones, J. A., Wang, E. Q., Chen, L., ... & Liu, H. (2016). The impact of Parkinson's disease on the cortical mechanisms that support auditory–motor integration for voice control. *Human brain mapping*, *37*(12), 4248-4261.

Knolle, F., Schwartze, M., Schröger, E., & Kotz, S. A. (2019). Auditory predictions and prediction errors in response to self-initiated vowels. *Frontiers in Neuroscience*, *13*, 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. *Journal of communication disorders*, *31*(6), 471-483.

Metz, O. (1952). Threshold of reflex contractions of muscles of middle ear and recruitment of loudness. *AMA Archives of Otolaryngology, 55*(5), 536-543.

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. *Journal of Neuroscience*, *33*(41), 16110-16116.

Numminen, J., & Curio, G. (1999). Differential effects of overt, covert and replayed speech on vowel-evoked responses of the human auditory cortex. *Neuroscience letters*, *272*(1), 29-32.

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

Toyomura, A., Miyashiro, D., Kuriki, S., & Sowman, P. F. (2020). Speech-induced suppression for delayed auditory feedback in adults who do and do not stutter. *Frontiers in Human Neuroscience*, *14*, 150.

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. *Cerebral Cortex*, *25*(6), 1576-1586.

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

Abstract

Determining if suspected speech 'abnormalities' in bilingual children are due to bilingual language acquisition or due to a speech sound disorder is a challenging task for speech-language pathologists (SLPs). This study aims to investigate how the productions on non-word imitation (NWI) of English-Dutch bilingual children differ from other speech tasks, both in direct comparison and relative to norm data.
77 typically developing Australian English-Dutch bilingual children ages 4 to 12 years, participated in this study. All children completed the Dutch test battery, the Computer Articulation Instrument (CAI). Data on language exposure were collected through parent/caregiver questionnaires.
The English-Dutch bilingual children scored lower than the norm data on picture naming and consistency task but not on non-word imitation and MRR tasks, confirming these tasks as the most language-neutral. A detailed phonological error analysis indicates that VOT, fricatives, and vowels need attention of SLPs assessing English-Dutch bilingual children.

Keywords: bilingual speech sound acquisition, non-word imitation, bilingual speech sound disorders, Dutch-English bilinguals, CAI

1. Introduction

There is always at least one bilingual child on the caseload of a speech-language pathologist (SLP) and the number of bilingual children has only grown over time (Wei Qin Teoh & McAllister, 2018). This poses a major challenge for SLPs, as often it is hard to distinguish whether the errors presented in bilingual children are due to cross-linguistic effects or an underlying disorder.

Several reasons contribute to this challenge. First, bilingual speech-sound acquisition is less researched and less understood compared to speech sound acquisition in monolingual children (Hambly et al., 2013). Bilingual speech sound development varies from that of monolingual children. There can be evidence of negative or positive transfer, and factors like language dominance play an important role in the pattern of development in bilingual children (Fabiano, 2023; Hambly et al., 2013). Secondly, the assessment procedures for the bilingual population are lacking in terms of norms and applicability. In a survey of 128 Australian SLPs, they preferred the use of informal assessment rather than standardized measures while evaluating the speech and language of multilingual children (Williams & McLeod, 2012). These challenges often lead to speech sound disorders in bilingual children being either under-referred or overrepresented, and this could have a lasting impact on the social and educational outcomes of the children (Hambly et al., 2013).

Therefore, the literature places a high emphasis on the need for more assessment tools, training, and research to improve assessment practices for bilingual children (Kohnert, 2010; Stow & Dodd, 2005). Existing assessment procedures for distinguishing speech sound errors typically rely on naming tasks evaluated on measures like percent consonant correct (PCC) and phonological error pattern analysis (Ortiz, 2021; Schwob et al., 2021). Studies show that on such tasks, bilingual children perform less compared to monolingual children. For example, a recent study reported higher rates of consonant substitutions and distortions, phonetic errors and phonological processes in Turkish–Dutch and Moroccan Arabic–Dutch bilingual children when compared to monolingual Dutch children (Alighieri et al., 2020).

Some studies have used non-word repetition to distinguish between typical errors and speech-sound disorders. Non-word repetition has the advantage of bypassing the lexical system and linguistic input. It has the potential for the diagnosis of speech sound disorder in multilingual children (dos Santos & Ferré, 2016; Ortiz, 2021).

In a nutshell, there is a need for more research to understand the speech-sound acquisition of bilingual children, as several factors can act as barriers that prevent the identification of speech sound disorders in these children. Additionally, there is a need for establishing reference data for typical bilingual speech sound development that SLPs could use while assessing bilingual children. For this purpose, the current study investigated speech sound development of Australian English–Dutch bilingual children. First, we evaluated different tasks for typically developing Australian-English–Dutch bilingual children and comparing their performance across these tasks. Secondly, we evaluated the error patterns in bilingual children and compared these to their monolingual peers.

2. Method

2.1. Participants

77 Australian English-Dutch bilingual children ranging between 4 and 12 years of age (M = 7.96, SD = 2.40; 43 girls, 34 boys) from the Dutch school in Sydney participated in this study. All children featured typical development and attended a regular Australian school. Additionally, they attended the Dutch school for two hours each week and they had 2 hours homework pertaining to the Dutch language. Written consent was sought from all children and their parents or caregivers prior to the study. Data on language exposure was collected through parent/caregiver questionnaires (a concise version of the anamnesis for multilingual children; Siméa, 2014). The parental reports indicated that 65% of the children spoke Dutch at home more than half of the time, 20% spoke a combination of English and Dutch at home more than half of the time, 9% spoke English at home more than half of the time, and 6% spoke

a combination of Dutch and another language at home more than half of the time.

2.2. Procedure

All children completed the Dutch standardized test battery *Computer Articulation Instrument* (CAI; Maassen et al., 2019), which includes the tasks picture naming, nonword imitation, consistency of word and nonword repetition, and diadochokinesis (maximum repetition rate). Picture naming (PN) covers the whole chain of the speech production process, from preverbal visual-conceptual processing to lemma access, word-form retrieval, phonological encoding, motor planning, –programming, and –execution. In the case of nonword imitation (NWI), lexical representations are not available, and the speaker must rely on either the phonological decoding and encoding system or the auditory-to-motor-planning pathway to produce the target utterance. In the consistency task then (word– and nonword repetition; WR and NWR), the child is asked to repeat words and nonwords in a sequence of five. This task assesses the variability of the produced speech output (percentage [non]-word-forms; PWF), which taps into motor planning and –programming and stability of the phonological representation (retrieved or constructed) of the (non)word form. Diadochokinesis (DDK) finally, tasks the children with repeating sequences (e.g., patakapataka…) as fast as possible (maximum repetition rate; MRR) and is thought to mainly reflect motor planning, programming, and execution. More information on the construction, reliability, validation, and norming of the CAI is available in (Diepeveen et al., 2019; van Haaften et al., 2019, 2021).

Data were collected with the CAI by student SLP's working in pairs using a computer or laptop, which automatically stored the acoustic signal on the hard disk. The children were seated in front of a microphone and wore open-back headphones to provide a good sound level of the automated instructions. The recordings were transcribed using broad phonetic transcription and analyzed on the computer by the student-SLPs according to the CAI examiner's manual (Maassen et al., 2019). All student SLPs were trained in the transcription and analysis protocol. Following the CAI psychometric evaluation guidelines (van Haaften et al., 2019), the training included practicing and evaluating the transcription and other analyses with two practice-examples of children with a Speech Sound Disorder. The transcriptions of the CAI of all children in this study were checked between the student-SLPs and differences were discussed. Consensus transcriptions were finally checked by the second author.

The data were collected at the Dutch school in Sydney. All children were given ample time to rest and play between tasks. Additionally, the Intelligibility in Context Scale-Netherlands (ICS-NL) questionnaire was completed by one of the children's parents/caregivers.

2.3. Data processing & analysis

Group-level quantitative phonological error analyses compared performance across tasks. Additionally, qualitative error analyses investigated the error patterns in terms of phonological processes. The CAI is normed for children up to 7 years old, but in this case also used with older children. For the analysis, the children were therefore split in two age groups: 4-7 years old (< 7 years; $n = 29$) and 7-12 years old (≥ 7 years; $n = 48$). The two age groups did not differ in gender distribution [X^2 (1, 77) = .320, $p = .57$].

For the younger children, the raw scores were transformed into z-scores per CAI norm age-group – to control for speech development to be able to compare the different variables with each other relative to the norm data for Dutch which is based on (predominantly) monolingual children. For the older children, the raw scores were transformed into z-scores based on the oldest CAI norm age-group (6;6–6;11 y;m). Note that this comes down to a linear transformation (as all datapoints are relative to the same reference data) and that the subsequent analysis is equivalent to an analysis of the raw data. It should further be noted that the data of the older children consequently are less reliable to compare with the norm scores of the CAI and may overestimate their performance relative to their age. However, in the present study it can be justified to present the results as z-scores –for a number of reasons. For the majority of outcome measures, the norm data show a ceiling effect for the 6-7 year age groups (van Haaften et al., 2019, 2021). This ceiling effect occurs for the main tasks of interest in the current study, PN, NWI, and consistency; DDK is the exception. In addition, our analysis focuses on between-task, within-subject differences. The analysis of the data from the older group is limited to a within-group comparison between tasks and does not include an analysis of age-related trends relative to norm data. Further, presenting the results as z-scores provides context necessary for interpreting between-task differences and occurrence of phonological processes.

Possible associations between tasks were examined through a correlational analysis, with the raw data (across age-groups). Correlations were calculated by means of Pearson's r. An α-value of .05 was used for all analyses.

The outcome of the picture naming and nonword imitation tasks will be analyzed using percent consonants corrects in the syllable initial position (PCCI) and using percent vowels correct (PVC). For the word and non-word consistency task percentage of different word forms were calculated. For the maximum repetition rates, the mean number of syllables uttered per second was calculated, subsequently averaged over all DDK subtasks (Mean MRR).

3. Results

Multivariate analyses of variance did not reveal any differences based on gender (multivariate nor univariate tests). Therefore, gender was not included as a factor or covariate in the final analyses.

Correlation analysis showed a significant positive association between the performance on all tasks and age (**Table 1**).

Table 1: *Correlation (Pearson's r) between all the tasks (raw scores) with age.*

	Age	PN - PCCI	PN - PVC	NWI - PCCI	NWI - PVC	WR	NWR
PN - PCCI	.50**	--					
PN - PVC	.33**	.62**	--				
NWI - PCCI	.51**	.54**	.37**	--			
NWI - PVC	.40**	.56**	.45**	.56**	--		
WR	.27*	.31**	.28*	.23*	.26*	--	
NWR	.48**	.28*	.19	.58**	.52**	.26*	--
Mean MRR	.61**	.48**	.11	.46**	.37**	.17	.31**

*. Correlation is significant at the 0.05 level (2-tailed);
**. Correlation is significant at the 0.01 level (2-tailed).

Results from one-sample t-tests comparing the group of 4-7 year-olds revealed they performed lower than the norm on the picture naming (PN-PCCI [$t(28)$=-7.124, p<.001]; PN-PVC [$t(28)$=-5.397, p<.001]) and the consistency tasks (WR [$t(28)$=-3.157, p=.004]; NWR [$t(28)$=-5.500, p<.001]). On nonword imitation (NWI-PCCI; NWI-PVC) and DDK (MRR), the scores on NWI and MRR were age appropriate (**Figure 1**).

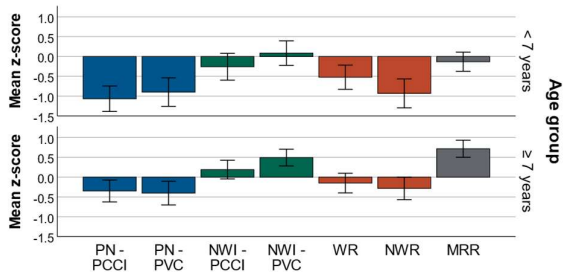


Figure 1: *Mean z-scores of Picture naming (PN), Nonword imitation (NWI), Nonword repetition (NWR) and Maximum Repetition Rate (MRR) tasks, broken down by age-group.*

Repeated measures ANOVA's were conducted to analyze within-subject effects for both age groups separately. For participants under 7 years, the analysis revealed a significant main effect of outcome measure [$F(6,22)$=19.792, p<.001]. The same pattern was observed for participants 7 years old or older [$F(6,42)$=15.483, p<.001).

Pairwise comparisons were conducted to assess differences between tasks. For both age groups, results indicated significant differences between both the two PN outcome measures (PCCI and PVC) on the one hand, and NWI-PCCI, NWI-PVC and Mean MRR on the other (all p's <.01). A same pattern was observed for the two consistency tasks (WR & NWR) apart from WR compared to NWI-PCCI, which did not reach statistical significance (all other p's <.05). No differences were found between PN-PCCI and PN-PVC nor between WR and NWR while NWI-PCCI and NWI-PVC were significantly different for both groups (both p's<.05).

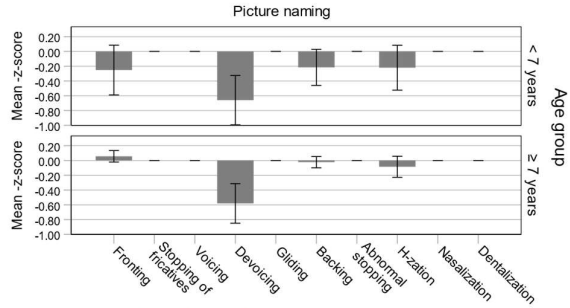


Figure 2. *Phonological processes (z-scores) in Picture Naming (PN) task.*

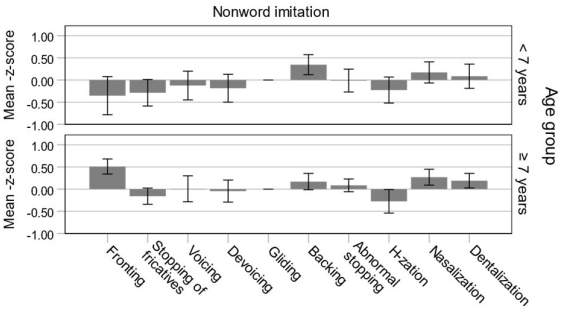


Figure 3. *Phonological processes (z-scores) in Non-Word Imitation (NWI) task.*

The observed phonological processes in the PN and NWI tasks are presented in **Figures 2 and 3**, respectively. The results indicate different patterns of phonological processes between the two tasks, which are consistently observed in both age groups. In PN, observed processes mainly comprise devoicing, fronting, backing, and h-zation while in NWI the observed processes seem to a cover broader range with predominance of fronting, stopping, and h-zation.

4. Discussion

The present study set out to investigate task-related differences in speech and motor-speech assessment in typically developing Australian English–Dutch bilingual children that were divided in two age-groups (4–7-year-olds and 7–12-year-olds). First, the results of the correlational analysis showed significant correlations between age and performance on all outcome measures, demonstrating their validity for assessing speech/speech-motor development in bilingual children.

The younger group was evaluated relative to norm data from Dutch monolingual typically developing children. For both groups, we then evaluated differences in performance between tasks, as well as patterns of phonological processes. The results indicated that the group of 4–7-year-old English-Dutch bilingual children scored lower compared to the norm data on percent consonants correct and percent vowels correct in picture naming (PN) as well as on consistency of word and nonword repetition (WR & NWR). While the children showed a small delay on these speech tasks, the scores on non-word imitation (NWI) and maximum repetition rate (DDK) were age appropriate. This pattern was confirmed by the pattern of between task pairwise comparisons which indicated significantly better performance on these two tasks compared with PN and consistency (WR & NWR). The group of 7–12-year-olds showed the same pattern. These results illustrate that use of a variety of tasks is necessary to understand the speech sound development in bilingual children. In particular, these findings corroborate earlier studies that have identified NWI as relatively language-neutral speech task. NWI is thought to bypass language processes involved in speech production, as opposed to picture naming, which would also activate lexical processing (dos Santos & Ferré, 2016; Ortiz, 2021). In addition, performance on the DDK task was like that on NWI. As such, NWI and DDK have large potential for the diagnosis of speech sound disorder in multilingual children.

The results of the phonological process analysis revealed differences in the occurrence of between PN and NWI, which were consistent across age groups. A particularly striking difference was observed for devoicing, which occurred often in PN but not in NWI. An acoustic measure such as voice onset time (VOT) might explain these devoicing errors. The VOTs of voiced vs. unvoiced plosives in Dutch don't match up with those in English. In Dutch, voiced stops are prevoiced (negative VOT) and unvoiced stops have a VOT of about zero, while in English, voiced plosives have a VOT of about zero and their unvoiced cognates are aspirated (positive VOT). Upon closer inspection, during picture naming the children tended to produce stop contrasts using English VOT's. However, these were all transcribed by the native Dutch student SLPs as voiceless stops as in Dutch both map onto the voiceless cognate. The transcriber's own phonology influences the perception of the produced phoneme, which in this case when the child produced the voiced plosive with English VOT, caused the sound to be perceived by the transcriber as its Dutch voiceless

cognate and hence resulted in the perception of a devoicing error. Interestingly, this pattern of excessive devoicing was seen only in picture naming and not in the non-word imitation task. Apparently, the children were able to perceive segments as pre-voiced and produce them accordingly when presented with an auditory model during non-word imitation. When they had to produce these sounds implicitly, as in a picture naming task, the children resorted to the use of English VOT for the plosive resulting in devoicing errors. Similar patterns have been reported in previous studies which observe crosslinguistic interaction in VOT in bilingual children (Stoehr et al., 2018).

Similarly, the patterns of processes like stopping of fricatives or h-zation might be explained by differences between the two languages in their fricatives' spectral center of gravity (COG). Typical COGs of fricatives in Dutch are different from those of their English counterparts, for example the COG of the Dutch /s/ lies in-between that of the English /s/-/ʃ/ contrast. Similar to the pattern for VOT, a pattern was observed in which fricatives productions in PN with English COG (spectral center of gravity) map onto a Dutch phoneme for the Dutch transcriber, while (failed) attempts to match the specific COG in NWI occasionally resulted in stopping or h-zation.

Finally, the results on the consistency tasks (5 consecutive repetitions of words and nonwords) showed occurrences of a unique pattern of increased transfer of English features with each subsequent repetition (e.g., telefoon: /teləfoːn/ > [tɛləfoːn] > [tɛləfon] > [tæləfon] > [thæləfon]). This pattern was observed in both similar ("olifant"/ˈɔo̯.liˌfant/-"elephant"/ˈɛlɪfənt/) and unsimilar words ("paraplu"/paːraːˈply/- "umbrella" /ʌmˈbrɛlə/) between English and Dutch. The memory trace of the acoustic model appears to fade with each repetition and the task thus slowly becomes a delayed imitation task. Subsequently the children seem to resort to use of English phonology and motor plans instead of the Dutch ones. Taken together, these results suggest that there is interference of English phonology and a loss of readily available phonological representations including motor goals for Dutch speech sounds.

5. Conclusion

The findings of the present study corroborate earlier findings of the nonword imitation task as being the most language neutral speech assessment task, and therefore of crucial importance as a diagnostic task to help identify whether speech sound development in bilingual children is age appropriate. Additionally, we believe that the use and comparison of different tasks is of particular importance as this can inform us about the underlying processes contributing to speech output. The present results showed differences in patterns of speech sound accuracy and phonological processes between PN and NWI as well as interesting patterns in the consistency of speech sound production.

We interpret these results to have two important implications. First, the speech errors in typically developing bilingual children have different origins depending on the task. While they appear to be due to the interference of English phonology in their Dutch productions in picture naming, they appear to be due to attempting to match a specific auditory model in non-word imitation. Second, the "ear of the beholder" appears to play an important role in the assessments. The transcriber's own phonological system influences the way the productions of the children are being perceived, meaning SLP's need to be aware of their bias and resort to, e.g., acoustic measurements when assessing phonetic or phonemic inventories and phonological processes. For the assessment of English-Dutch bilingual children, SLPs particularly need to pay close attention to voicing errors, fricative and vowel productions, and h-zation.

6. Acknowledgements

The authors express gratitude to the two anonymous reviewers of our conference proposal for their constructive comments. We also extend our thanks to all the children and parents/caregivers who took part in the study.

7. References

Alighieri, C., D'haeseleer, E., Daelman, J., Lancker, F. V., Laperre, M., Kissel, I., & Lierde, K. V. (2020). Articulation skills in bilingual children with a migration background: A comparison between bilingual Turkish-Dutch, Arabic-Dutch and monolingual Dutch children. *Journal of Communication Disorders*, *87*, 105993.

Diepeveen, S., van Haaften, L., Terband, H., de Swart, B., & Maassen, B. (2019). A standardized protocol for Maximum Repetition Rate assessment in children. *Folia Phoniatrica et Logopaedica*, *71*, 238–250.

dos Santos, C., & Ferré, S. (2016). A Nonword Repetition Task to Assess Bilingual Children's Phonology. *Language Acquisition*, *25*.

Fabiano, L. (2023). Introduction to the Research Symposium on Bilingualism Forum. *Journal of Speech, Language, and Hearing Research*, *66*(12), 4673–4677.

Hambly, H., Wren, Y., McLeod, S., & Roulstone, S. (2013). The influence of bilingualism on speech production: A systematic review. *International Journal of Language & Communication Disorders*, *48*(1), 1–24.

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of Communication Disorders*, *43*(6), 456–473.

Maassen, B., van Haaften, L., Diepeveen, S., Terband, H., van den Engel-Hoeck, L., Veenker, Th., de Swart, B. (2019). "Computer Articulatie Instrument: Handleiding". *Boom Uitgevers*.

Ortiz, J. A. (2021). Using Nonword Repetition to Identify Language Impairment in Bilingual Children: A Meta-Analysis of Diagnostic Accuracy. *American Journal of Speech-Language Pathology*, *30*(5), 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. *Journal of Speech, Language, and Hearing Research*, *64*(9), 3578–3593.

Siméa, S. (2014). Indicatiecriteria: Auditief en/of communicatief beperkte leerlingen. *Retrieved from www.Simea.Nl/Dossiers/Passend-Onderwijs/Brochures-Po/Simea-Brochure-Indicatiecriteria-Juni-2014.pdf.*

Stow, C., & Dodd, B. (2005). A survey of bilingual children referred for investigation of communication disorders: A comparison with monolingual children referred in one area in England. *Journal of Multilingual Communication Disorders*, *3*(1), 1–23.

van Haaften, L., Diepeveen, S., Terband, H., de Swart, B., van Den Engel-Hoek, L., & Maassen, B. (2021). Maximum repetition rate in a large cross-sectional sample of typically developing Dutch-speaking children. *International Journal of Speech-Language Pathology*, *23*(5), 508–518.

van Haaften, L., Diepeveen, S., van den Engel-Hoek, L., Jonker, M., de Swart, B., & Maassen, B. (2019). The psychometric evaluation of a speech production test battery for children: The reliability and validity of the Computer Articulation Instrument. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2141–2170.

Wei Qin Teoh, C. B., & McAllister, S. (2018). Bilingual assessment practices: Challenges faced by speech-language pathologists working with a predominantly bilingual population. *Speech, Language and Hearing*, *21*(1), 10–21.

Williams, C. J., & McLeod, S. (2012). Speech-language pathologists' assessment and intervention practices with multilingual children. *International Journal of Speech-Language Pathology*, *14*(3), 292–305.

Spatiotemporal variation of tongue dorsum characterizes the voicing contrast of American English bilabial coda obstruents

Daejin Kim

Department of Linguistics, University of New Mexico, USA

daejinkim@unm.edu

Abstract

This study explores the articulation of the tongue dorsum (TD) in voicing contrasts of bilabial coda obstruents in American English (AE) using the ultrasound tongue imaging method. TD, where the tongue curvature begins, is known to constrict during the bilabial closure to create a sufficient amount of aerodynamic pressure indicating voicing. Also, its movement direction and duration should correlate to the acoustic properties of coda voicing due to the linkage to the laryngeal structure creating these attributes. The qualitative and quantitative analysis of ultrasound recordings of seven AE speakers found that TD is more likely to constrict with /b/ with the longer moving distance and duration, but the movement direction is mostly affected by the phonetic quality of the preceding Vs. TD constrictions of C2 and V occur earlier with /b/, resulting greater lingual-laryngeal gestural aggregation, making V more like V and C2 more overlapped with V. This study suggests that the intricate temporal adjustment of the lingual and laryngeal articulation implicates the AE bilabial voicing consonantal contrasts.

Keywords: ultrasound tongue imaging, tongue dorsum, bilabial tongue constriction, American English, coda voicing contrasts

1. Introduction

The lip opening and closing are, without a doubt, the primary articulatory characteristics of bilabial stop consonants, but several studies (Ahn, 2018; Fuchs *et al*., 2004; Lindblom *et al.*, 2002; Svirsky *et al*., 1997; Vazquez-Alvarez & Hewlett, 2007) have found and considered that bilabial obstruents should be conditioned with the tongue body lowering, related to the increase of aerodynamic pressure within the oral cavity (Stevens & Hanson, 2010). The tongue moves towards the pharyngeal wall with the voiced obstruents to indicate the increased aerodynamic pressure during the bilabial closure. It has been termed as a '**trough effect**' or '**closure-related tongue perturbation**'. This may be either due to the speakers' active control of laryngeal and supralaryngeal articulators or as an articulatory consequence of the passive deformation from the relaxation of supraglottal muscles.

Among various articulatory attributes of the AE bilabial voicing obstruents, the lower and posterior parts of the tongue are particularly relevant to this effect because of the proximity to the laryngeal structure and the relatively wide oropharyngeal space suitable to hold the air coming out of the lung. **The tongue dorsum (TD)**[i], located at the very back of the tongue, where the tongue curvature begins, and/or where the tongue surface is under the velopharyngeal port when resting and under the velum when articulating velar obstruents, has been increasingly understood as an important articulatory property that distinguishes the voicing properties of bilabial obstruents across languages, which also correlates to the acoustic properties (*f0* and vowel duration) of those contrasts. Varying degrees of TD displacement and duration should play an important role in contrasting the voicing and laryngeal properties of bilabial obstruents and, therefore, should contrast the consonantal qualities in speech. The precise nature of this phonetic behavior, however, has been unknown with several conflicting results.

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). Several studies have explored the underlying mechanism of the coda voicing contrasts due to its obscure implementation of acoustic and articulatory attributes. This study argues that the voicing contrast of C2 should be evidenced in the spatiotemporal movement of TD, following the evidence of TD movement during the closure of onset obstruents from previous studies. Svirsky *et al.* (1997) reported that English speakers showed higher intra-oral air pressure during the consonantal closure with /ɑbɑ/ than with /ɑpɑ/, and the smaller tongue displacement with /p/ is due to the speakers' intentional relaxation of tongue muscles. The larger tongue displacement with /b/ was argued due to the active manipulation of tongue muscles. Vazquez-Alvarez & Hewlett (2007) reported that the displacement of the tongue exists during the oral closure for bilabial stop consonants in British English, but the direction of the tongue displacement and the tongue contour shape differences highly vary across contexts and by individual speakers. Ahn (2018) also reported that the voiced obstruents /b, d, g/ were articulated with the advanced TR and fronted tongue body positions than the voiceless counterparts /p, t, k/ in Brazilian Portuguese and English, even though AE voiced consonants are considered technically voiceless due to its short voice onset time (VOT). However, no previous studies reported the relevance between the tongue displacement and the voicing contrasts of bilabial obstruents at the coda of a syllable in AE.

This secondary supralaryngeal movement would also occur during the acoustic V duration to signal coda 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 laryngeal structure and decrease the tension of the vocal folds, is more likely to occur with low Vs and C2 /b/, resulting in lower *f0*, while TD raising, which may raise the larynx and increase the tension of the vocal folds, is more likely to occur with high Vs with C2 /p/, resulting in higher *f0* (Ohala, 1978). Coretta (2020) reported that the longer vowel duration of the following stop consonant in the CVCV context corresponds to greater TR advancement at the offset of the preceding vowel (the end of vocal fold vibration) in Italian and Polish. If AE follows similar articulatory programming similar to the finding, the current study also argues that greater TR advancement is positively correlated with longer vowel duration which may contribute to signaling the coda voicing contrast in AE. However, little evidence for such a pattern has been reported in AE.

This study hypothesizes that C2 /b/ should be associated with more frequent TD raising or lowering with longer and larger moving distance and duration and greater intergestural timing between articulatory landmarks of C2 and

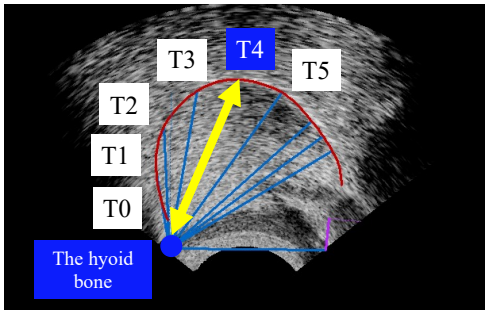
V than C2 /p/. This would imply that TD movement distance and duration from the onset to the target and from the target to the offset may contribute to the acoustic characteristics of bilabial C2s, signaled by acoustic V duration.

2. Methods

Audio-synchronized ultrasound tongue images (frame rate = 54-60 Hz; transducer radius = 20mm, depth = 75-90 mm, Angle = 103.2°) were collected from seven native speakers of AE (three males and four females (mean age = 21.4) from Colorado, Texas, and New Mexico in the USA) using the Telemed Echoblaster 128 and the Articulate Advance Assistant software (Articulate Instruments, 2023). All recordings were conducted at a sound-attenuated booth in the Linguistics Graduate Lab at the Department of Linguistics. Each speaker produced six monosyllabic target words (/hVC2/; V = /i, u, ɑ/, C2 = /p, b/; *heap*, *heab*, *hoop*, *hoob*, *hop*, *hob*) in sentence-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 focus (Q: Did you write ***god*** on the {paper, note, board, letter}? A: No. I wrote a ***heap*** (/hip/) on the {paper, note, board, letter}) prominence six times. 504 produced tokens were collected.

Tongue contours were estimated consecutively over time using DeepLabCut™ (Wrench & Balch-Tomes, 2022). It automatically marks darker edges under the brighter reflections from the tongue surface (T4). All estimated contours before, after, and during the target words were visually inspected and manually corrected if those contours needed to be adjusted or estimated incorrectly. Tokens of which tongue contours were invisible in recorded images and unable to be corrected were excluded from the analysis. After annotation and analysis, 454 produced tokens were analyzed. TD movement landmarks were annotated as distance and time between the hyoid bone and the tongue surface (T4) (**Figure 1**) when it starts (ONSET), reaches its target (TARGET), and ends its movement (OFFSET) for V and C2. Dynamic characteristics of TD were then quantified by calculating the moving distance and duration of TD from the onset to the target and from the target to the offset.

Figure 1 A sample illustration of the TD movement distance measured from the hyoid bone to the tongue surface.



Since not all tokens have visually apparent TD movement, tokens were annotated as 'Yes' if TD movements were visually evident or 'No' if not apparent in the images. 'Yes' tokens were classified again based on TD onset-to-peak distance as either 'Raising' (≥ 0) or 'Lowering' (< 0). To estimate which factors contribute to the occurrence and presence and absence of TD movements on bilabial C2s, the data was qualitatively analyzed by building conditional inference trees (CITs) (Levshina, 2020; Schweinberger, 2023) using the *partykit* (Hothorn *et al*., 2023) and *ggparty* (Borkovec *et al*., 2019) packages in *R*.

All extracted ([x, y] coordinates) tongue contours at the TD targets of bilabial C2 were rotated and scaled from the estimated vector from the shadow of the hyoid bone to the shadow of the mandible to normalize the difference in the size and shape of the tongue as well as the mandible of individual speakers. The tongue contour shape differences were assessed using the generalized additive mixed effects model using *mgcv* (Wood, 2017) and *itsadug* (van Rij *et al*., 2022) packages.

TD movement differences among consonants were assessed with (i) the tongue shape contour differences when it maximally contrasting C2s and Vs, (ii) onset-to-target and target-to-offset distance (.mm) and duration (.ms), (iii) intergestural timing (.ms) between V and C2 onsets, targets, and offsets, and (iv) timing of V and C2 onsets, targets, and offsets from the end of acoustic vowel duration to gauge the laryngeal-lingual articulatory coordination in timing. The statistical significances of movement variables were statistically tested as a function of *f0* peaks, vowel duration, C2s (/p/ and /b/), vowels (/i/, /u/, /ɑ/), and focus prominence types (BF and CF) 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*. Readers should note that only statistically significant variables were included in the models.

3. Results

3.1. The occurrences and moving direction of TD perturbation for bilabial C2s

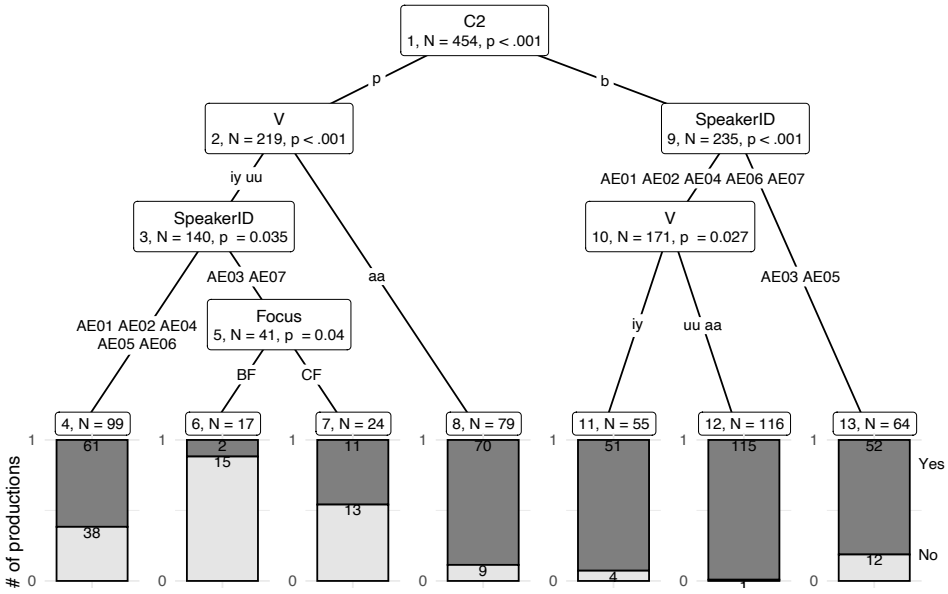
A CIT model (**Figure 2**) estimating the factors that affect the presence and absence of TD movement (accuracy rate = 0.79) found that C2 /b/ (probability of 'Yes = 0.93) is more likely to have TD constriction compared to C2 /p/ (0.66). Among /p/ tokens, /ɑ/ (0.89) is more likely to have TD constriction during V than /i/ and /u/ (0.56). Two speakers (AE03 and AE07; #5) seem to move TB more often with CF (#7 = 0.46) with the /i/ vowel than with BF (#6 = 0.12). Consonant voicing (/p/ and /b/) is the most important variable in the occurrences of TB movements on bilabial C2s. **Table 1** includes the calculations of the probabilities of 'Yes' on each node.

In terms of TD movement direction (a separate model; no figures were included.), the 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). This indicates that the TD bilabial constriction is likely affected by the quality of preceding Vs.

Table 1 The probability of tokens with TD constrictions on each node of the CIT model.

Node	Descriptions	Yes	No	Probability of 'Yes'
1	Total (N = 454)	362	92	0.80
2	C2 = /p/	144	75	0.66
3	V = /i/ and /u/	74	66	0.53
4	AE01, ..., AE06	61	38	0.62
5	AE03, AE07	13	28	0.32
6	Focus = BF	2	15	0.12
7	Focus = CF	11	13	0.46
8	V = /ɑ/	70	9	0.89
9	C2 = /b/	218	17	0.93
10	AE01, ..., AE07	166	5	0.97
11	V = /i/	51	4	0.93
12	V = /u/ and /ɑ/	115	1	0.99
13	AE03, AE05	52	12	0.81

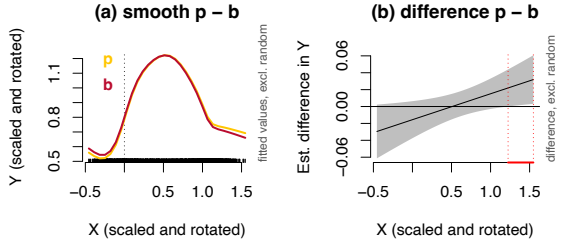
Figure 2 The Conditional Inference Tree model estimating the occurrence of TD movements during the acoustic V duration ('iy' = /i/, 'uu' = /u/, 'aa' = /ɑ/).



3.2. The tongue contour shape differences contrasting the voicing of bilabial C2s

The generalized additive mixed-effects model (**Figure 3**) estimated that the difference in tongue contour shape between /p/ and /b/ at the coda of a syllable is statistically significant (C2 = /p/: *edf.* (estimated degree of freedom) = 8.11; F = 120.00, $p < .001$; C2 = /b/: *edf.* = 9.11; F = 124.95, $p < .05$); however, the model comparison with the current model and the model without the C2 variable found that the model difference was statistically insignificant. It only estimates the effect of the quality of the preceding Vs; the tongue at the maximally constricting position for the C2 is more fronted and raised with /i/ than /with /ɑ/, while the tongue is more retracted and lowered with /ɑ/ compared to /u/ (V = /i/: *edf.* = 8.11; F = 120.00, $p < .001$; V = /u/: *edf.* = 170.99; F = 124.95, $p < .001$; V = /ɑ/: *edf.* = 14.39; F = 218.24, $p < .001$). Therefore, no positional and morphological tongue shape differences exist between /p/ and /b/ at the coda of a syllable, which contradicts the previous findings.

Figure 3 The GAMM result showing (a) the estimates of the tongue contour smooths of bilabial C2s' TD target positions and (b) the difference between the two levels of the C2 variable.



3.3. The movement characteristics of TD for bilabial C2s

Regarding TD movements' characteristics, the models estimated that C2 /b/ has a *longer* onset-to-target distance of TD (only when raising) (β (estimate of the difference between /p/ and /b/) = -0.6*) **(a)** and a *longer* onset-to-target duration (β = -6.4**) **(b)**, compared to C2 /p/ (i.e., /p/ < /b/). The target-to-offset distance (β = -0.4) (c) and duration differences of TD (β = -3.1) (d) were not statistically different between C2s (i.e., /p/ = /b/). Also, the models estimated no statistically significant effect of the voicing quality of bilabial C2s in TD and TD onset-to-target and target-to-offset distance and duration of the preceding Vs.

Regarding intergestural timing (**Figure 5**), TD starts to move simultaneously for V away from C2 onsets between C2 and V (**(a)** β = -2.4) (i.e., /p/ = /b/). /b/ has the *earlier* V targets (**(b)** β = -10.7*) and offsets of TD (**(c)** β = -12.6***) than /p/ (i.e., /p/ < /b/), resulting in less coarticulation between V and C2 in timing. /b/ was articulated with *later* C2 target and offset of TD than /p/, resulting in articulatory expansion away from V. About the relative timing of TD landmarks away from the acoustic end of V, /b/ has earlier onset, target, and offset of TD for both V (**(d)** V onset – V end: β = 39.7*; V target – V end: β = 40.3***; V offset – V end: β = -37.5***) (i.e., **/Vp/ > /Vb/**) and C2 (**(e)** C2 onset – V end: β = 35.9***; C2 target – V end: β = 29.8***; C2 offset – V end: β = -24.9***) compared to /p/ (i.e., **/p/ > /b/**). All TD movements of /b/ occur earlier than those of /p/, resulting in more temporal overlap with the laryngeal property of V (i.e., V duration) and the supralaryngeal property of C2 (TD movements for C2).

Figure 4 The point range plots (mean (point) and 1 stand deviation (ranges)) showing the estimates of (a) onset-to-target distance, (b) duration, (c) target-to-offset distance, and (d) duration of TD.

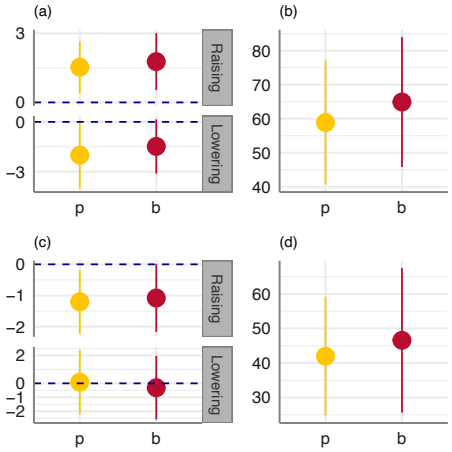
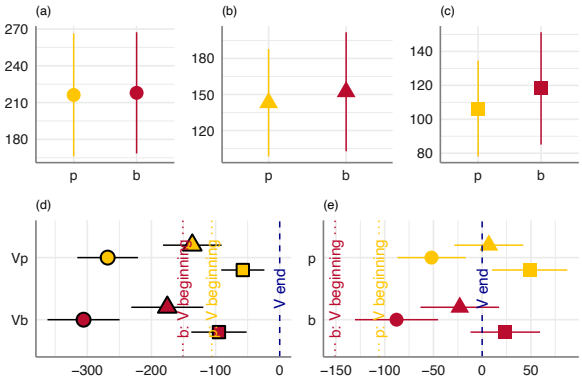
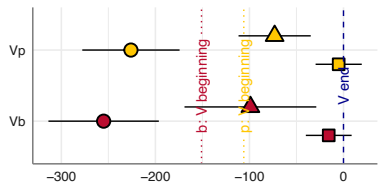


Figure 5 The point range plots showing the estimates of the intergestural timing between (a) C2 and V onsets, (b) targets, and (c) offsets. The relative timing of TD landmarks away from the acoustic end of V (blue dashed lines) of (d) V and (e) C2.



This study also estimated the relative timing of produced tokens with no visually apparent TD movements for bilabial C2s ('No' tokens). /b/ has *earlier* TD onsets and targets of the preceding V from the acoustic V end (V onset – V end: β = 29.02*; V target – V end: β = 25.48*) (i.e., **/Vp/ > /Vb/**) (a), but the offset difference between /Vp/ and /Vb/ was estimated statistically insignificant (V offset – V end: β = 8.36) (i.e., **/Vp/ = /Vb/**) (**Figure 6**). 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.

Figure 6 The relative timing of TD landmarks away from the acoustic end of V (blue dashed lines) of V gestures of tokens with no apparent TD movements for bilabial C2.



4. Discussion and conclusion

Overall, the voicing quality (i.e., /b/) on the bilabial coda obstruent seems to be articulated with TD constriction more frequently by either raising or lowering TD than /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 these qualitative variations as a result of the phonetic underspecification (Choi, 1995; Honorof, 1999; Keating, 1988) of TD constriction for bilabial C2s. TD constriction is more likely to occur to indicate voicing, but its occurrences are subject to individual variations, and the movement direction is phonetically affected by the quality of the preceding Vs. Also, some speakers were found to constrict TD more frequently with CF than with BF; this can be interpreted as TD constriction as a phonetic enhancement.

These TD actions during the acoustic V duration, therefore, enhance the phonetic quality of the preceding V. 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 measured during the acoustic V duration. Instead, it can be interpreted that the longer movement distance and duration with /b/ may be positively correlated with the longer V duration with the following Vs.

The acoustic consequence of voicing from the laryngeal articulation (i.e., lower *f0*) was not directly associated with the articulation of TD. Also, the articulatory characteristics of /p/ and /b/ still seem unclear with mixed results. Instead, voicing coda contrast is more apparently accompanied by gauging intergestural timing between TD landmarks and acoustic events. The 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. A similar timing pattern was also estimated among produced tokens with visually inapparent (or no) TD movements in ultrasound images. This study, therefore, argues that these varying timing patterns during V contrasting consonantal voicing can be considered a part of the phonetic grammar for AE speakers.

Taking all results together, this study suggests that the voicing contrasts of bilabial coda obstruents /p/ and /b/ are conditioned with the intricate temporal timing control of the laryngeal property of V and the supralaryngeal properties of bilabial C2. In conclusion, TD movement should be regarded as an important articulatory correlate indicating the voicing property of bilabial C2 obstruents in AE.

5. Acknowledgments

I thank Sarah Lease and Seth Wyatt for recording stimulus sentences, and Ryan Smith and the research participants who wanted to be anonymous for sharing their voices and tongues for this study. This study was supported by the Graduate Student Success Scholarship from the College of Arts and Sciences at the University of New Mexico.

6. 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.

Articulate Instruments. (2023). Articulate Assistant Advanced (AAA) (221.02.0) [Computer software]. Articulate Instruments.

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

Borkovec, M., Madin, N., Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., & Yutani, H. (2019). ggparty: "ggplot" Visualizations for the "partykit" Package (1.0.0) [Computer software].

Choi, J. D. (1995). An acoustic-phonetic underspecification account of Marshallese vowel allophony. *Journal of Phonetics*, 23(3), 323–347.

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.

Fuchs, S., Hoole, P., Brunner, J., & Inoue, M. (2004). The trough effect— An aerodynamic phenomenon? From *Sound to Sense*.

Honorof, D. N. (1999). *Articulatory Gestures and Spanish Nasal Assimilation* [Dissertation]. Yale University.

Hothorn, T., Seibold, H., & Zeileis, A. (2023). partykit: A Toolkit for Recursive Partytioning (1.2-20) [Computer software]. https://cran.r-project.org/web/packages/partykit/index.html

Keating, P. A. (1988). The phonology-phonetics interface. In F. J. Newmeyer (Ed.), The Cambridge linguistic survey, vol I: Linguistic theory. Cambridge University Press.

Levshina, N. (2020). Conditional Inference Trees and Random Forests. In M. Paquot & S. Th. Gries (Eds.), A Practical Handbook of Corpus Linguistics (pp. 611–643). Springer.

Lindblom, B., Sussman, H. M., Modarresi, G., & Burlingame, E. (2002). The Trough Effect: Implications for Speech Motor Programming. *Phonetica*, *59*(4), 245–262.

Maddieson, I. (1997). Phonetic Universals. In W. J. Hardcastle & J. Laver (Eds.), The Handbook of Phonetic Science (1st edition) (1st edition, pp. 619–639). Blackwells.

Munhall, K., & Löfqvist, A. (1992). Gestural aggregation in speech: Laryngeal gestures. *Journal of Phonetics*, *20*(1), 111–126. https://doi.org/10.1016/S0095-4470(19)30242-6

Ohala, J. J. (1978). Production of Tone. In V. A. Fromkin (Ed.), Tone: A Linguistic Survey (pp. 5–39). Elsevier.

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 12(1), 6.

Schweinberger, M. (2023). Tree-based models in R. https://slcladal.github.io/tree.html

Stevens, K. N., & Hanson, H. M. (2010). Articulatory–Acoustic Relations as the Basis of Distinctive Contrasts. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), The Handbook of Phonetic Sciences (2nd ed., pp. 424–453). Wiley.

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.

van Rij, J., Wieling, M., Baayen, R. H., & Rijn, H. van. (2022). *itsadug*: Interpreting Time Series and Autocorrelated Data Using GAMMs (2.4.1) [Computer software].

Vazquez-Alvarez, Y., & Hewlett, N. (2007). The 'Trough Effect': An Ultrasound Study. *Phonetica*, *64*(2–3), 105–121.

Wood, S. N. (2017). Generalized Additive Models: An Introduction with R (2nd ed.). Chapman and Hall/CRC.

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.

[i] Several studies have termed this part of the tongue as 'tongue body' interchangeably, typically 4-5cm away from the tongue tip (or the tongue tip magnetic sensor for electromagnetic articulography studies) (Rebernik et al., 2021). TD in this study is located further behind the back of the tongue. It is also possible to term this part of the tongue as the tongue root.

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, Austria

yaqian.huang@oeaw.ac.at

Abstract

How subtypes of creaky voice such as vocal fry and period doubling can be classified according to their acoustic as well as phonatory correlates is not entirely clear. This study explores the distinctions of the above-mentioned creaky voice types as compared to modal voice, using machine classification methods to investigate the importance of source and filter characteristics represented by acoustic and electroglottographic (EGG) measures. Tokens of vocal fry, period doubling, and modal voice were visually identified in a scripted Mandarin corpus using EGG as these non-modal voice qualities were found abundantly across Mandarin tones. To control for the multicollinearity and overfitting issues, an l1 regularization (Lasso) was used to fit the multinomial logistic regression. Random forest models were also used to predict these voicing types and compared with the logistic models. Adding the EGG measures largely improved all model performances, both supported by the separable clusters shown by explorative visualization and the macro average precision and recall scores. The most important measures according to the random forest models were f0, H1-H2, H1, SoE, H2, and HNR (0-500Hz), as well as the duration of the decontacting phase and contact quotient of the glottal pulse. Implications between human perception and phonatory measures are discussed.

Keywords: voice quality, machine learning, vocal fry, period doubling, electroglottography (EGG)

1. 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 and/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 the two 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.

Vocal fry is often used across languages as a prosodic element which tends to occur utterance-finally, or plays a role in different social interactions and attitudes (Davidson 2021). Period doubling is oftentimes referred to as "diplophonia" in speech pathology (Schreibweiss-Merin & Terrio 1986) and is used in several singing styles including throat singing and Sardinian singing (Bailly *et al*. 2010). More importantly, vocal fry and period doubling were commonly found allophonic to Mandarin tones (Yu 2010), and appear to have different linguistic distributions driven by a focus-marking position (vocal fry) or utterance edges (period doubling) (Huang 2023). This study thus uses continuous read speech in Mandarin from multiple speakers. 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.

2. Methods

2.1. Materials

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 ȵi3 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. **Figure 1** shows representative samples of vocal fry, period doubling, and modal voice in EGG and audio waveforms. On an EGG waveform, the portion from the most positive slope above the valley to the most negative slope after the peak is commonly calculated as the glottal contacting phase, indicated by the blue square, whereas the portion from the most negative slope to the most positive slope around the valley is the glottal open

phase, indicated by the red square. In particular, vocal fry is characterized by low f0 and high glottal constriction shown by the fatter-look glottal pulse, and modal voice has nearly identical pulses with balanced open and contacting phases. Period doubling is characterized by alternating pulses between high-low amplitudes and/or long-short periods.

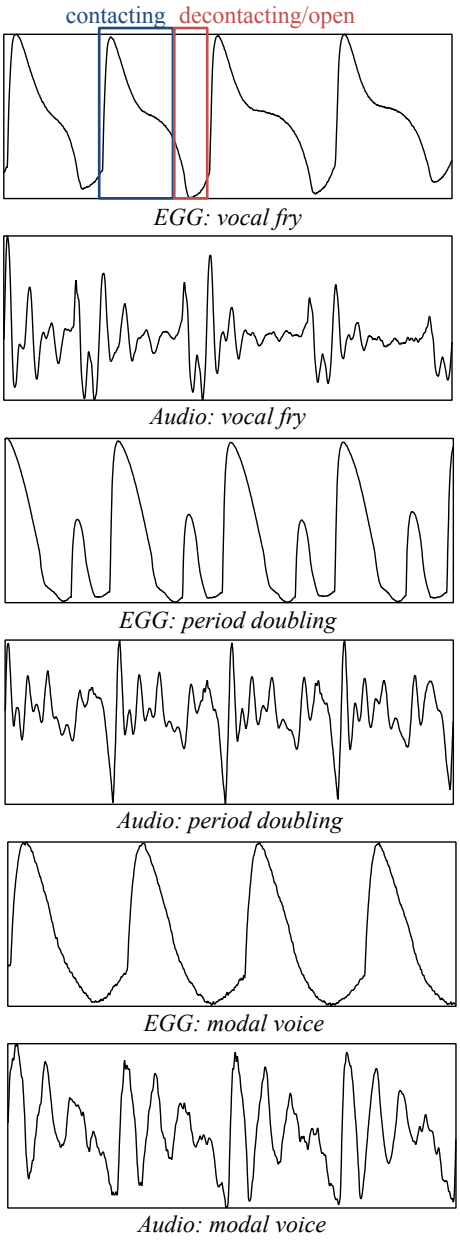


Figure 1: *Representative samples of period doubling, vocal fry, and modal voice tokens.*

2.2. Measures and analyses

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 formant-and-bandwidth corrected harmonics and spectral tilts, harmonics-to-noise ratio, cepstral peak prominence, subharmonic-to-harmonic ratio, formants and bandwidths, and energy measures with epoch (the instant of significant excitation/glottal closure); 11 EGG measures included different proportions of contacting and decontacting durations, contact quotient based on different calculation methods (CQ_H, CQ_HT, CQ_PM; Tehrani 2009, also see reviews in Herbst 2020), speed quotient (ratio, defined as the ratio of the duration of the contacting phase to the duration of the decontacting phase), 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.
Table 1 shows the overall distributions of the three voicing types across men and women in the two datasets. 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. R packages *Rtsne*, *glmnet*, and *randomForest* were used.

Table 1: *Distributions of vocal fry, period doubling, and modal voice in women and men in the acoustic and the acoustic+EGG dataset.*

Acoustic	Vocal fry	Period doubling	Modal voice
Women	482	2354	1175
Men	156	943	428
Total	638	3297	1603
Acoustic +EGG	Vocal fry	Period doubling	Modal voice
Women	154	324	187
Men	25	91	137
Total	179	415	324

To assess the modal performance, overall accuracy, precision, and recall scores were used. Macro average was used because of the imbalance in the counts of vocal fry, period doubling, and modal voice tokens. The formulas of accuracy, macro average of precision and recall, and precision and recall are given in equations (1-5).

$$Accuracy = True\ Positive/All\ Tokens \quad (1)$$

$$Macro\ Average\ Precision = Average(Precision(pd)+Precision(fry)+Precision(modal)) \quad (2)$$

$$Macro\ Average\ Recall = Average(Recall(pd)+Recall(fry)+Recall(modal)) \quad (3)$$

$$Precision = True\ Positive/(True\ Positive+False\ Positive) \quad (4)$$

$$Recall = True\ Positive/(True\ Positive+False\ Negative) \quad (5)$$

3. Results

3.1. Exploration using t-SNE clustering

T-SNE clusters based on voicing types are shown in **Figure 2**. Both 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. This corresponds to findings in Huang 2022 that glottal pulses seem to alternate between different degrees of glottal constriction, resulting in shared characteristics between both creaky and modal voice. Adding EGG measures helps separate the subtypes of creaky voice, so that three clusters corresponding to each of the voicing types are observed.

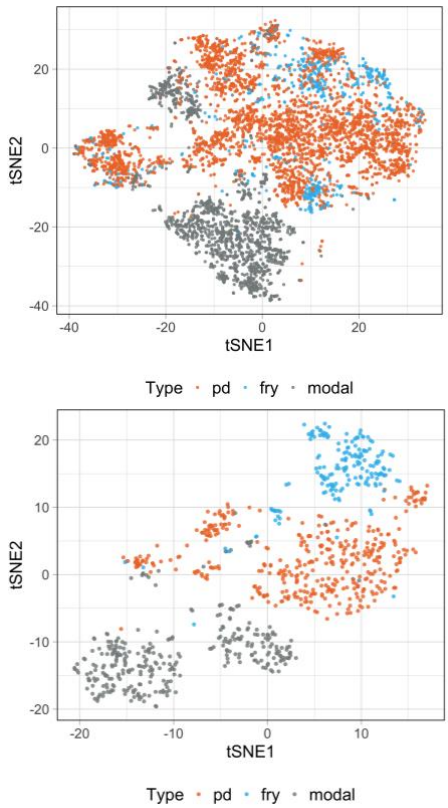


Figure 2: *t-SNE clustering of tokens of period doubling (orange), vocal fry (blue), and modal voice (gray) using only the acoustic (left) and both acoustic and articulatory measures (right). The datasets have different sizes.*

3.2. Classification results

Table 2 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 2: *Summary of overall accuracy and macro average precision and recall scores using different machine learning methods.*

Acoustic	Logistic Lasso regression	Random forest
Accuracy	0.9112	0.9312
Macro avg. precision	0.8749	0.9098
Macro avg. recall	0.8137	0.8529
Acoustic + EGG	Logistic Lasso regression	Random forest
Accuracy	0.9837	0.9967
Macro avg. precision	0.9840	0.9975
Macro avg. recall	0.9784	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(0-500Hz) are the most important; further, the duration between 10% and 90% of the glottal decontacting phase (SQ4-SQ3) and contact quotient (CQ) of the glottal pulse are the most important EGG measures among all. In both datasets, H1*–H2*, H1*, SoE remain most important. **Figure 3** and **Figure 4** show the top 15 ranks of important acoustic and articulatory features in the two datasets.

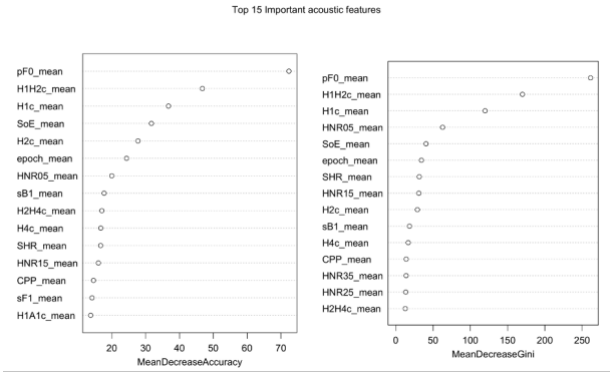


Figure 3: *Top 15 important acoustic features in the training set of the random forest model of the acoustic dataset.*

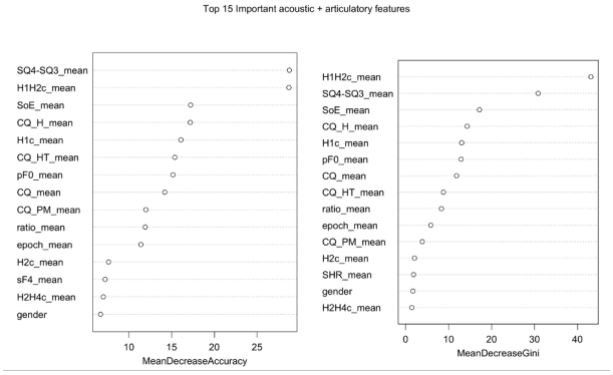


Figure 4: *Top 15 important acoustic and articulatory features in the training set of the random forest model of the combined dataset.*

4. Discussion and conclusion

This study used logistic Lasso regression and random forest models to investigate how subtypes of creaky voice including vocal fry and period doubling can be differentiated among each other and with modal voice, using acoustic features as well as phonatory measures obtained by EGG. The models using acoustic features alone already show reasonable separation with a recall score of 85%, whereas the models using a larger set with both acoustic and articulatory features more effectively distinguish period doubling, vocal fry, and modal voice from each other with a higher score around 99% for accuracy, precision, and recall using random forest. The most important acoustic measures are f0, H1*–H2*, and H1*, and the most important EGG features are the duration between the 10% and 90% 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.

Because vocal fry has a disproportionately fewer number of tokens, a common pitfall across such algorithms is the misclassification of vocal fry, compared to other voicing types, especially for the acoustic dataset. More balanced datasets are thus desirable for a machine learning problem, but it may also be worth in a follow-up study to closely examine the files that are easily misclassified as another type to reduce potential conflation, or motivate theoretical questions such as how acoustically distinct we could establish two voice categories, how to allocate importance of articulatory characteristics and acoustic attributes when distinguishing different voice types, etc.

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. Yet, given that period doubling and vocal fry are distinguishable from each other and modal voice (Gerratt & Kreiman 2001), it is possible that the perceptual product is not only acoustics, but a combination of both vocal fold vibration and vocal tract resonances. Moreover, such implications for the lack of salience of voice types for listeners may be biased by the selection of the cues investigated (different voice qualities can introduce different f0 dynamics), by the manual labeling strategy, by the speakers' task, and by the language under examination.

To remedy the indirect relationship between actual perception and source dynamics, in future work, one could inspect the characteristics of voice qualities shown by the source, as modeled using the EGG signal, and the filter, reflected in the audio signal, for the classification and prediction of the labels of the voice categories. For example, we could devise a bi-directional classifier trained on EGG or audio signals and tested on their counterparts, namely, using EGG waveform to predict acoustic parameters of a particular voice type and vice versa. The relation between the source and filter could be evaluated by machine-learning approaches which take articulatory or acoustic features of pulses to form a selection criterion and in turn test on the corresponding counterpart chunks to see how well the EGG and audio signals can predict each other in terms of the occurrence of period doubling or vocal fry. Questions to ask further are, what aspects of vocal fold articulation during a particular voicing type lead to changes in the audio signal? How to predict the changes in perceived resonance as a function of changes in the source? Then we could study how changes in vowel quality interact with the overall phonatory pattern. A caveat here might be that the EGG signal, though a model of the source, will not be the same as the aerodynamic source being filtered by the vocal tract to establish the more direct mapping and correspondence.

5. Acknowledgements

The author would like to thank Eran Mukamel, Ed Vul, Will Styler, Sarah Creel, Marc Garellek, Eva Reinisch, Leon Bergen, and Hong Zhang for their comments. This material is based upon work supported by the NSF under Grant No. BCS-2141433.

6. References

Bailly, L., Henrich, N., and Pelorson, X. (2010). Vocal fold and ventricular fold vibration in period-doubling phonation: Physiological description and aerodynamic modeling. *The Journal of the Acoustical Society of America, 127*(5), 3212–3222.

Davidson, L. (2021). The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages. *WIREs Cognitive Science, 12*(3), e1547.

Garellek, M. (2019). The phonetics of voice. In W. F. Katz & P. F. Assmann (Eds.), The Routledge Handbook of Phonetics. Abingdon, Oxon; New York, NY: Routledge. 75-106.

Gerratt, B. R., & Kreiman, J. (2001). Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, *29*(4), 365-381.

Hedelin, P. and Huber, D. (1990). Pitch period determination of aperiodic speech signals. *International Conference on Acoustics, Speech, and Signal Processing*, 361–364.

Huang, Y. (2022). Articulatory properties of period-doubled voice in Mandarin. *Proc. Speech Prosody 2022*, 545-549.

Huang, Y. (2023). *Phonetics of period doubling*. University of California, San Diego.

Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. *Proceedings of the 18th ICPhS*, 2–7.

Kreiman, J., Gerratt, B. R., Precoda, K., and Berke, G. S. (1993). Perception of supraperiodic voices. *The Journal of the Acoustical Society of America, 93*(4), 2337–2337.

Redi, L. and Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics, 29*(4), 407–429.

Schreibweiss-Merin, D. & Terrio, L. M. (1986). Acoustic analysis of diplophonia: A case study. *Perceptual and motor skills, 63*(2), 755–765.

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. *Proceedings of the ICPhS XVII*, 1846-1849.

Tehrani, H. (2009). EGGWorks: a program for automated analysis of EGG signals. Computer program.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9*(11), 2579-2605.

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

Abstract

When engaging in a conversation, interlocutors frequently accommodate to each other in their speech patterns and co-speech movements. However, only a small number of studies have investigated both domains in a multimodal approach. An additional challenge for studies is accounting for information structure, which not only influences the production of speech and co-speech motion in a speaker but also affects the patterns of accommodation between speakers. Due to the high complexity of the required experimental design, it has not yet been comprehensively studied whether speakers accommodate to each other in their strategies of encoding information structure. This paper present a methodological approach for capturing multimodal focus marking patterns in dyads, which allows to address this research question. We introduce DiCE*, a cooperative game to elicit lexically and prosodically controlled data in German, and present details of the experimental setup involving dual EMA, audio, and video.*

Keywords: electromagnetic articulography, dual EMA, accommodation, multimodality, focus structure.

1. Introduction

Previous research has demonstrated that speakers frequently accommodate to their interlocutors' speech patterns and speech-accompanying movements. There is evidence for convergence of, e.g., head motion (Hale *et al.* 2020), manual gestures (Mol *et al.* 2012), and postural sway (Shockley, Santana, & Fowler 2003). Interlocutors may also accommodate in terms of intonation (Babel & Bulatov 2012), speaking rate and phrasing (Cummins 2002), as well as acoustic properties of vowels and consonants (Pardo *et al.* 2012; Nielsen 2011). Leveraging recent technological advancements, a limited number of studies have used dual electromagnetic articulography (dual EMA) to elucidate accommodation in supra-laryngeal speech kinematics, reporting results for jaw, lip, and tongue movements (Lee *et al.* 2018; Mukherjee *et al.* 2018; Tiede & Mooshammer 2013). However, only a few studies have integrated the multiple modalities of accommodation within a single experimental setting yet (but cf. Duran & Fusaroli 2017; Louwerse *et al.* 2012; Oben & Brône 2016). To our knowledge, only one study has analysed multimodal accommodation using dual EMA, with results presented for only one dyad (Tiede *et al.* 2010).

One factor that markedly shapes a speaker's production of speech and co-speech movements is information structure (Ladd 2008; Wagner, Malisz & Kopp 2014). Depending on, e.g., the focus structure of an utterance, a particular word may be produced with a larger F0 rise, a more distinct tongue articulation, and/or a more pronounced head nod. Typically, we deal with a high amount of speaker-specific variability in patterns of focus marking. This complicates the design of studies on accommodation because comparing words that occur under different structural circumstances may confound the results. Furthermore, it has been shown that words are more sensitive to interpersonal accommodation when they occur in prosodically salient positions (Lee *et al.* 2018). This underscores the potential of including controlled information structure in experimental designs, particularly in those targeting accommodation. What remains an open area for investigation is whether these patterns of marking information structure (or more specifically, focus) are themselves subject to interpersonal accommodation. We aim to address this question in the future through an analysis of the recorded data set described below.

In this paper, we present a comprehensive methodological approach to capturing multimodal accommodation across various acoustic, visual, and kinematic levels of speech production. With this, we hope to contribute valuable insights for future studies sharing similar research goals. We introduce *DiCE (Dialogic Collecting Expedition)*, a cooperative card game in German that provides a natural context to elicit speech material controlled for segmental context of target words and focus structure of utterances. We provide practical information on the experimental setup with dual EMA, audio, and video. The method allows to capture multimodal accommodation of focus encoding patterns within dyads. We have successfully applied it in recordings of 15 dyads of German native speakers.

2. Methods

The complete game material for *DiCE* is publicly available for future use at https://osf.io/9fmqh/.

2.1. Technical set-up and procedure

EMA and audio recordings are conducted using two 3D electromagnetic articulographs (Carstens AG501 and AG501 Twin) and two head-mounted condenser microphones (MicroMic C544 L, connected to a MicroMic MPA V L phantom adapter). The technical setup is schematically illustrated in Figure 1. Each of the two articulographs is connected to a unique SyBox (Carstens SyBox2), which are interlinked and connected to an interface (Tascam US4x4). Additionally, each articulograph is linked to a unique recording laptop, and these laptops are interconnected (via a router that does not access the internet) to allow for data transfer. The two microphones are plugged into the same interface as the two articulographs, which enables a temporal synchronisation of the signal streams. The interface is connected to the recording laptop associated with EMA1, where each recording sweep is initiated. The EMA signal is simultaneously recorded on both articulographs at 1250Hz and then downsampled to 250Hz. The audio is recorded at 48kHz with a bit depth of 16.

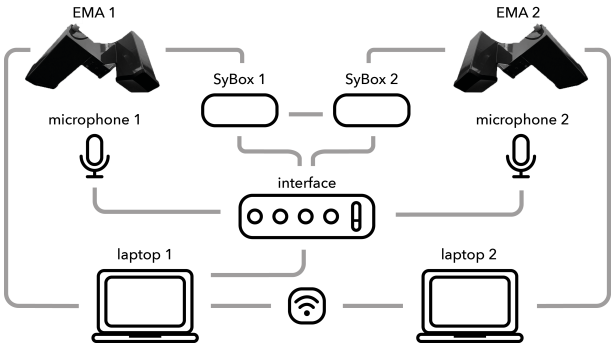


Figure 1: *Recording setup for EMA and audio.*

EMA sensors are attached to both speakers on the torso (both shoulders, chest, and spine), the head (forehead, eyebrows, and behind both ears), and the articulators (lower jaw, upper and lower lip, both corners of the mouth, tongue tip, and tongue dorsum), as illustrated in Figure 2.

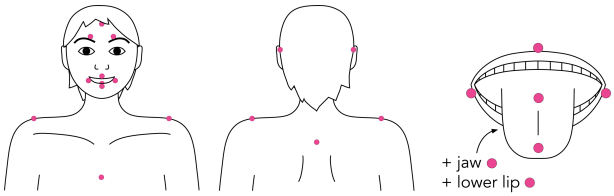


Figure 2: *Illustration of EMA sensor placements.*

In addition to EMA and audio recordings, videos are captured using three cameras strategically positioned in the room: Two cameras are placed on the table between the two articulographs, each directed at one participant (2x GoPro Hero9 Black), and one camera is positioned to capture both participants together from the side (Panasonic HC-V520). The videos are recorded at a resolution of 1080p, a frame rate of 50fps, and a constant shutter speed of 1/100s. Post-hoc synchronization of the videos to the EMA and audio recordings is achieved using the auditory signal from a clapperboard at the beginning of each recording sweep.

Each recording lasts between three and four hours, including preparations and breaks. The recordings consist of two main parts: one solo condition per participant and one dialogue condition with both participants (cf. Figure 3). In the solo condition, each speaker is recorded individually in a simplified digital version of the experimental game. After both speakers have completed their solo condition, they are introduced to each other and cooperatively play the experimental game in the dialogue condition.

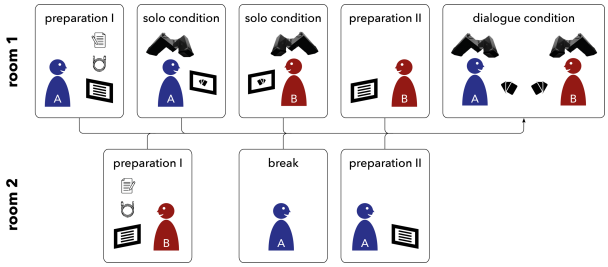


Figure 3: *Scheme of recording procedure with two participants (A and B) in two rooms. Prep. I includes informed consent, attachment of EMA sensors and instructions for the solo condition, prep. II involves a break and instructions for the dialogue condition.*

2.2. Speech material

The speech material produced in the experiment is highly controlled both lexically and prosodically. To allow for an analysis of supra-laryngeal articulation using EMA, eight carefully chosen target words are included, comprising four objects and four cities (cf. Table 1). Their lexically stressed penultimate syllable, which occurs in a controlled segmental context, can be used for analyses. As objects, we selected existing words in German. As cities, we selected two words for existing cities (*Medina*, *Manila*), one word for an existing city borrowed from Italian (*Milano*), and one pseudoword (*Benali*). Participants in our recordings encountered no difficulties in producing these words nor did they report any unease.

Table 1: *Target words.*

objects		**cities**	
Bohne *bean*	[ˈboːnə]	Medina	[meˈdiːna]
Mode *fashion*	[ˈmoːdə]	Manila	[maˈniːla]
Vase *vase*	[ˈvaːzə]	Milano	[miˈlaːno]
Made *maggot*	[ˈmaːdə]	Benali	[beˈnaːli]

These target words are embedded in consistent question-answer sets in German, which the two participants produce. Each set contains two questions and two answers (cf. Table 2). Speakers are instructed to consistently adhere to the lexical structure of the carrier sentence and only replace the object and/or city in the utterance. This approach ensures that the corpus includes a substantial amount of lexically consistent speech material.

Table 2: *Exemplary question-answer set.*

Q1	Habe ich die Bohne aus Medina auf der Hand? *Am I holding the bean from Medina?*
A1	Du hast die Mode aus Medina auf der Hand. *You are holding the fashion from Medina.*
Q2	Wo? *Where?*
A2	Da. *There.*

Particularly utterance A1 is of interest, since its information structure is controlled. Based on the preceding question Q1, it is produced with one of three possible focus structures: either (i) the object is in *corrective focus* and the city is in the *background*, as in the example in Table 2, (ii) the object is in the *background* and the city is in *corrective focus*, or (iii) both the object and city are in *corrective focus*.

2.3. Task

The speech material is obtained through a custom-designed card game called *DiCE (Dialogic Collecting Expedition)*. In the dialogue condition, the game is played cooperatively by the two participants, while in the solo condition, a simplified digital version is played by each participant separately in front of a screen. In the game narrative, the subjects assume the roles of collectors who have discovered a basement filled with valuable items. These items are the target words, i.e. the four objects from the four cities described above (cf. Table 1). They are represented by 16 playing cards (4 objects × 4 cities). The participants' objective in the game is to organise the cards based on the objects' values and origins. They succeed when they have

collectively arranged the cards into four piles in the middle of the table – one pile for each city of origin, with the four objects in ascending order from value one to four.

In the dialogue condition, each participant has three cards on a stand in front of them, which are positioned in a way that allows them to see only the other participant's cards, not their own. Through the question-answer sets (cf. Table 2), each participant aims to identify their own cards. One participant asks if they are holding a specific card, and their interlocutor replies. Then, they ask where, and their interlocutor replies and points to the intended card. When the participant finds a suitable card in their hand, they place it on the table, contributing to the incremental and cooperative ordering of the collection. A photo of the game in the dialogue condition is shown in Figure 4.

Figure 4: *Two participants sitting underneath articulographs, playing the dialogic game.*

Crucially, the task is complicated through constraints on the cards participants are allowed to mention in their responses. These constraints are designed to elicit the three focus structures for utterance A1 (cf. Table 2), as described before. They are introduced as strict communication rules within the community of collectors. Participants are prohibited from answering with a simple "yes" or "no" to their interlocutor's question Q1. Instead, they are required to refer to a different card than the one asked about, but always one genuinely present on their interlocutor's stand. This can be accomplished by substituting either the word for the object or the city mentioned in the question. For instance, if one participant asks if they have the vase from Milano, the other speaker can, e.g., refer to the *bean* from Milano, or to the vase from *Benali*, or, if no suitable card is present, to the *bean* from *Benali*. In this manner, speakers produce one of the three intended focus structures in the target utterance A1. The two participants take turns and have the freedom to choose which card to ask about and which one to refer to in their response, within the given rules. Penalty points are assigned in cases where they refer to a false card or fail to adhere to the specified lexical structures. In total, six rounds of the game are played in the dialogue condition. The number of question-answer sets produced in one round varies between dyads and rounds.

In the solo condition, where only one speaker is present in the room, they are seated in front of a screen and engage with a simplified digital version of the game (cf. Figure 5). In this setup, the participant exclusively responds to questions and does not initiate questions themselves. To present the stimuli, OpenSesame (Mathôt, Schreij & Theeuwes 2012) is used. For each question-answer set, three cards along with the question Q1 are displayed on the screen. The participant produces their response A1. Subsequently, the screen displays the question Q2, and the speaker points towards the intended card while producing their response A2. In total, each participant is prompted to produce 76 answers for both A1 and A2 (4 objects × 4 cities × 2 focus conditions × 2 renditions + 12 additional trials), with randomisation of trial order applied per participant.

Figure 5: *One participant sitting underneath the articulograph, playing the game in the solo condition*.

2.4. Corpus example

For our data set, 30 participants were recorded with the described methods at I*f*L Phonetics, University of Cologne, Germany. The participants ranged in age from 18 to 36 years (mean: 24.67, SD: 4.51) and had grown up in Germany with German as their native language. Six of the participants were bilingual, having at least one additional native language, but German was reported as the dominant language for all speakers. 17 of the subjects were female, 12 male and one non-binary. The participants were naive to the purpose of the experiment and did not possess advanced knowledge of phonetic sciences. While some participants mentioned the ability to speak a German dialect, they all spoke accent-free standard German during the experiment. For each recording, two participants were paired into a dyad with no constraints other than not being previously familiar with each other, resulting in a total of 15 dyads.

Following the data recording, several processing steps were required before the analysis of the data set. The video files were synchronised with the microphone-recorded audio files using the auditory signal from the clap, and they were trimmed to the same length using DaVinci Resolve. The audio files were transcribed with automatic speech recognition using OpenAI Whisper and manually checked for errors. Then, they were annotated based on the transcript using the Montreal Forced Aligner (McAuliffe *et al.* 2017), with manual corrections and the addition of further annotation tiers. EMA files were processed using the ema2wav converter (Buech *et al.* 2022).

An example of the multimodal data set that we recorded with the presented methods is illustrated in Figure 6, showcasing an exemplary question-answer set from the dialogue condition of one dyad. Four parameters (namely lip aperture, vertical tongue, head, and eyebrow motion) are selected from the extensive array of possible supra-laryngeal and co-speech kinematics, aiming to exemplify the nature of the recorded multimodal data.

3. Discussion and conclusion

We introduce a methodological approach for capturing multimodal accommodation, providing practical details on the technical setup, procedure, speech material, and task. Through the cooperative card game *DiCE (Dialogic Collecting Expedition)*, lexically and prosodically controlled German speech material is elicited within an engaging scenario, and is recorded with dual 3D electromagnetic articulography, audio, and video. The synchronisation of multiple data streams makes it possible to analyse a wide range of parameters in the auditory and visual modalities, potentially shedding new light on their spatiotemporal interrelations. Through this novel approach, it can be investigated in a fine-grained manner whether multimodal patterns of focus encoding are subject to interpersonal accommodation.

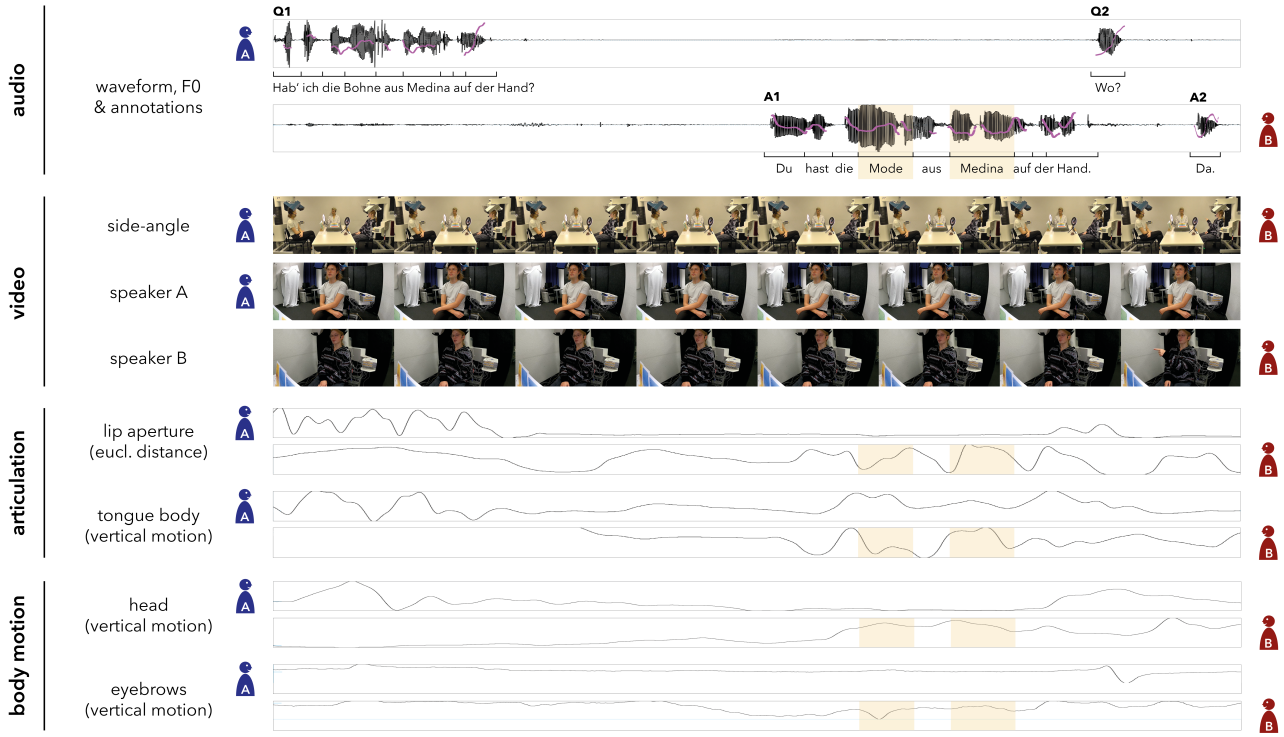


Figure 6: *Visualisation of recorded multimodal data for four selected parameters during one question-answer set by one dyad (speakers A and B). Based on the question Q1 (speaker A), the answer A1 (speaker B) has a focus structure with the target object in corrective focus and the target city in the background. Target words in A1 are marked by yellow rectangles.*

4. Acknowledgements

The authors would like to thank Theo Klinker, Tabea Thies, Katinka Wüllner, and Elisa Herbig for their help with the recordings. This work was supported by the German Research Foundation (DFG) as part of the SFB1252 "Prominence in Language" (Project-ID 281511265, project A04) and the a.r.t.e.s. Graduate School for the Humanities Cologne.

5. References

Babel, M., & Bulatov, D. (2012). The Role of Fundamental Frequency in Phonetic Accommodation. *Language and Speech*, 55(2), 231–248.

Buech, P., Roessig, S., Pagel, L., Hermes, A., & Mücke, D. (2022). ema2wav: doing articulation by Praat. *Proceedings of Interspeech, 18-22 September 2022*, *Incheon, Korea,* 1352–1356.

Cummins, F. (2002). On synchronous speech. *Acoustic Research Letters Online*, 3(1), 7–11.

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.

Hale, J., Ward, J. A., Buccheri, F., Oliver, D., & Hamilton, A. F. d. C. (2020). Are You on My Wavelength? Interpersonal Coordination in Dyadic Conversations. *Journal of Nonverbal Behavior*, 44, 63–83.

Ladd, R. D. (2008). *Intonational Phonology.* Cambridge Univ. Press.

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.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, 36, 1404–1426.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2), 314-324.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. *Proceedings of Interspeech, 20-24 August, Stockholm, Sweden*, 498–502.

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.

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.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, 39, 132–142.

Oben, B., & Brône, G. (2016). Explaining interactive alignment: A multimodal and multifactorial account. *Journal of Pragmatics*, 104, 32–51.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, 40, 190–197.

Shockley, K., Santana, M.-V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), 326–332.

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, Cancún, Mexico*, 4pSC10.

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*, 3aSCa3.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication,* 57, 209-232.

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

Abstract

The purpose of this study was to investigate chewing efficiency and oral sensory and motor developmental symptoms and habits in a clinical population of 210 Swedish children between 4 to 9 years old. The children were referred to Oral Motor Centre (OMC), Danderyd Hospital in Stockholm questioning oral- and/or speech motor disorders. Comparison was made with 77 typically developing children (TD). Chewing efficiency was performed with the Hue-Check© chewing gum test and analyzed with the outcome measure SDHue. Statistical analysis with multiple linear regression showed that children with oral developmental delay (OD) and motor speech disorders (MSD) chewed less efficient than the TD children. The children with OD, MSD and language disorders (PDL) also showed larger frequencies of oral sensory and motor symptoms and habits compared to the TD children. Impaired chewing and oral symptoms and habits can possibly be part of a larger symptom complex. Inability to chew efficiently can affect quality of life.

Keywords*: Developmental speech disorders; chewing efficiency; oral sensory symptoms.*

1. 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 (Newmeyer et al. 2009; Nijland et al. 2015). The children also often show 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). Due to the complexity and heterogeneity of symptoms differential diagnosis between subtypes can be 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, OD and PDL compared to children with typical development (TD) using the Hue-Check chewing gum test. Evaluation of oral sensory and motor symptoms and habits related to oral development was also explored.

2. 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 Swedish Ethical Review Authority. Informed consent was obtained from or on behalf of all participants.

Between January 2022 to October 2023, patients referred to OMC between 4 to 9 years with questions of oral- and speech motor dysfunctions were asked to participate in the study. 210 fulfilled the assessment (SG group; 147 boys, mean age = 6:2 years, range 4:0 - 9:6 and 63 girls, mean age = 5:8 years, range 4:0 - 9:3). Of these 210 patients 122 children agreed to participate in the chewing study (HSG group; 83 boys, mean age = 6:0 years, range 4:2 - 9:3 and 39 girls, mean age = 5:4, years, range 4:0 - 7:6). Data from a control group (CG group) was collected between March 2022 and March 2023. 77 (37 boys, mean age = 6:1, range 4:0 - 9:2 and 40 girls, mean age = 6:6, range 4:0 - 9:4).

The collection of data from the 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. 76 TD children agreed to participate in the chewing gum test. Oral

sensory and motor symptoms and habits were assessed through the anamnesis's information from a written questioner and from verbal interviews with the caregivers. Assessment of the Babkin reflex is included in and assessed by The Swedish translated version (not published) of the Verbal Motor Production Assessment for Children (VMPAC).

Chewing efficiency was assessed using the Hue-Check chewing gum test (a 2-colored chewing gum, one pink and one blue) and based on the coloring pattern of the two colors 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 122 specimens were photographed from both sides by the SLP clinician responsible for assessing the child. The photo samples of the chewing gums were administrated into a computer file and distributed to and assessed by a single calibrated operator, experienced with the SDHue analysis. The outcome measure SDHue is a color dispersion metric calculated by automated image analysis software.

Statistical analyses were conducted through SPSS using Shapiro-Wilk test, Independent T-test, Pearson Chi-Square test and multiple linear regression analyses, age separate and gender and age together as covariate.

Table 1: *Sample of entire study group (SG) and the children that chewed the HueCheck gum (HSG) presented in the diagnostic subgroups. MSD (Motor speech disorder): (Speech Motor Delay [ATYP]; Speech Motor Delay with CAS-features [ATYP+]; Childhood Apraxia of Speech [CAS]; Dysarthria [DYS]; Oral dyspraxia affecting speech [OA]. OD (Oral Motor developmental delay), with no MSD and PDL (Phonological Disorder Language): (Phonological Language Disorder [PLD]; Expressive Language Disorder [ELD] and Developmental Language Disorder [DLD]). Age 4 to 9 years.*

Subgroups Diagnoses	SG *n*	HSG *n*
MSD (all)	105	71
OD (all)	70	31
MSD-only	13	7
MSD+OD	40	27
MSD+OD+PDL	27	22
MSD+PDL	25	15
OD-only	33	14
OD+PDL	37	17
PDL-only	22	13

Table 2: *Results of Mean SDHue per diagnostic subgroup in age group 4 to 9. MSD (Motor speech disorder): (Speech Motor Delay [ATYP]; Speech Motor Delay with CAS-features [ATYP+]; Childhood Apraxia of Speech [CAS]; Dysarthria [DYS]; Oral dyspraxia affecting speech [OA]. OD (Oral Motor developmental delay), with no MSD and PDL (Phonological Disorder Language): (Phonological Language Disorder [PLD]; Expressive Language Disorder [ELD] and Developmental Language Disorder [DLD]) conducted with multiple linear regression analyses and gender and age together as covariate for adjusted p-values. All groups were compared to control. Standard Deviation (SD) and Beta is presented.*

Study Group	*n*	Mean SDHue	SD	P adjusted for gender and age	Beta
HCG	76	.48	.211		
MSD (all)	71	.62	.180	.009**	.202
OD (all)	31	.58	.191	.046*	.151
MSD-Only	7	.59	.208	.004**	.262
MSD+OD	27	.68	.194	<.001***	.371
MSD+OD +PDL	22	.57	.178	.324	.081
MSD+PDL	15	.58	.117	.530	.164
OD-Only	14	.60	.235	.373	.073
OD+PDL	17	.57	.151	.045*	.167
PDL-Only	13	.49	.188	.619	- .041

* = $p < .05$, ** = $p < .01$, *** = $p < .001$

3. Results

After assessment at OMC, 137 children in the SG group received an OD diagnose and 105 children an MSD diagnose: **(**Speech motor delay origin [ATYP]; Speech motor delay with CAS features [ATYP+]; Childhood Apraxia of Speech [CAS]; Dysarthria [DYS]; Oral dyspraxia affecting speech motor control [OA**)**. 111 children were prior assessment at OMC diagnosed with PDL: (Phonological language disorder [PLD]; Expressive language disorder [ELD] and Developmental language disorder [DLD]) as shown in Table 1. In the HSG group ($n = 122$), 7 children were diagnosed with MSD-only, 14 with OD-only and 13 with PDL-only. 27 had a combination of MSD and OD, 22 with MSD, OD and PDL, 15 with MSD and PDL and 17 with OD and PDL. (Table 1). 7 children are not included in the analyses of SDHue (5 not receiving a diagnosis and 2 children with ankyloglossia). The entire SG group was included in analyses of oral sensory and motor symptoms and habits. 13 children are not included in the analyses (9 not receiving a diagnosis, 3 children with ankyloglossia and one child with closed nasality diagnose).

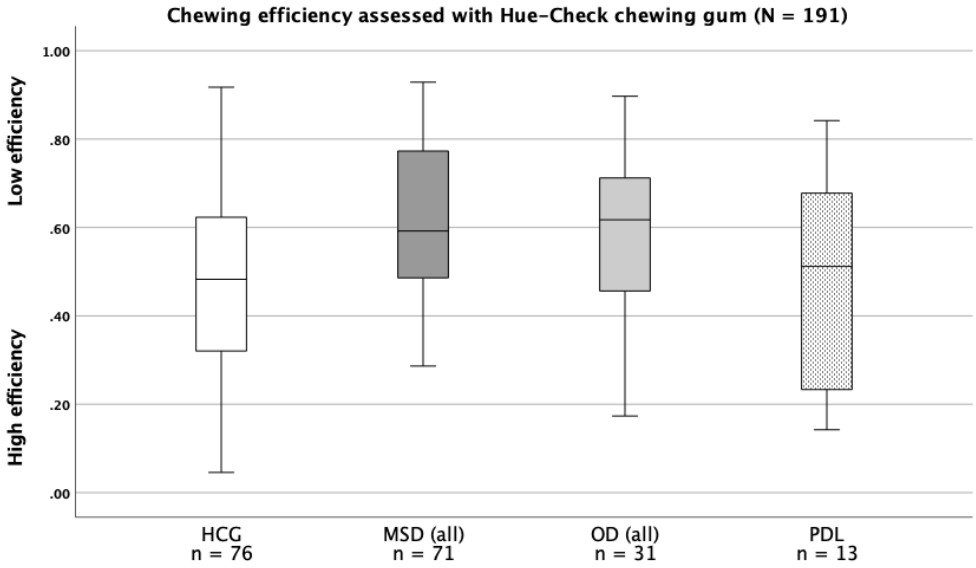


Figure 1: *Boxplot showing chewing efficiency measured through SDHue for the study group compared to the control group (HCG). The study group is divided in MSD (Motor speech disorder): (Speech Motor Delay [ATYP]; Speech Motor Delay with CAS-features [ATYP+]; Childhood Apraxia of Speech [CAS]; Dysarthria [DYS]; Oral dyspraxia affecting speech [OA]. OD (Oral Motor developmental delay), with no MSD and PDL (Phonological Disorder Language): Phonological Language Disorder [PLD]; Expressive Language Disorder [ELD] and Developmental Language Disorder [DLD]). Age 4 to 9 years.*

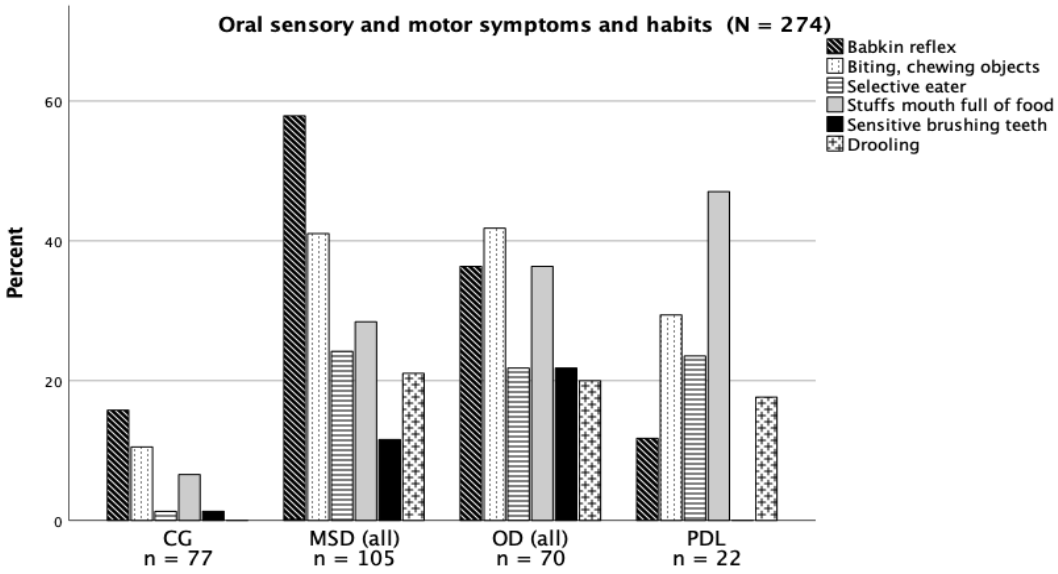


Figure 1: *Graph showing oral sensory and motor symptoms and habits in percent: Babkin reflex; Mouth Stimuli (chewing and biting on objects); Selective eater; Stuffs mouth full of food; Sensitive brushing teeth and Drooling. The entire Study group (SG) is compared to controls (CG) and divided in MSD (Motor speech disorder): (Speech Motor Delay [ATYP]; Speech Motor Delay with CAS-features [ATYP+]; Childhood Apraxia of Speech [CAS]; Dysarthria [DYS]; Oral dyspraxia affecting speech [OA]. OD (Oral Motor developmental delay), with no MSD and PDL (Phonological Disorder Language): Phonological Language Disorder [PLD]; Expressive Language Disorder [ELD] and Developmental Language Disorder [DLD]). Age 4 to 9 years.*

Descriptive statistics were calculated on the entire study group (SG; n = 210) for age, gender, and MSD diagnoses and were compared with the children that agreed to participate in the Hue-Check task (HSG; n = 122). Descriptive statistics showed satisfactory equivalence regarding age, gender, and diagnoses. Regarding the control group (CG; n = 77) there was satisfactory equivalence with SG and HSG regarding age but there was a difference in gender. Normal distribution analyses conducted with Shapiro-Wilk test showed satisfactory results in each diagnostic subgroup for both boys and girls. Further analyses of chewing efficiency between gender groups with Independent T-test showed no differences in chewing efficiency between groups (p=.85).

Statistical analyses showed that both the children with MSD-only and the children with a combination of MSD and OD chewed significantly less efficient compared to the controls (p=.004; p=<.001). The children diagnosed with OD-only did not chew significantly less efficient (p=.373). The entire MSD group (n = 72) as well as the OD group without MSD (n = 31) chewed significantly less effective than the TD group (p=.009; p=.046) shown in Table 2 and Figure 1. Comparison through Post Hoc analyses between the entire MSD and OD without MSD groups did not reveal a difference in chewing efficiency (p=.847).

Regarding oral sensory and motor symptoms and habits including *Babkin reflex*, *Mouth Stimuli (chewing and biting on objects), Selective eater*, *Stuffs mouth full of food, Sensitive brushing teeth* and *Drooling* children with MSD, OD and PDL showed higher frequencies of each symptom (p's<.001), apart from *Sensitive brushing teeth* (p=.008). Details of distribution of each symptom in percent is shown in Figure 2.

4. Discussion and conclusion

The present study investigated chewing efficiency and oral sensory and motor developmental symptoms and habits in a diverse clinical population of 210 4- to 9-year-olds. The 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, OD and PDL showed 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.

5. Acknowledgements

We are grateful to the children and caregivers from the study group and the control group approving to take part in this study. We also want to thank the speech language pathologist colleagues at OMC that helped to collect the data. A special and warm thank you to colleague Isabell Hammarlund.

We would like to thank Professor Martin Schimmel at University of Bern for all help with questions regarding the Hue-Check chewing gum and a warm thank you to Flavio for help with analyses of the chewing gum samples.

6. References

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.

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

Abstract

It has been reported that the time needed to produce a phoneme is highly variable, even for the same speaker, yet the causes of this variability have not been thoroughly studied. In this direction, the present work explored the effect of three factors on phoneme articulation time: attention, coarticulation and intentional speech rate (fast/slow). More precisely, the muscular activity of participants' lips was recorded while they produced the syllables /pa/ and /pu/, which made it possible to calculate each phoneme articulation time. In addition, this protocol was complemented with different tasks that made it possible to measure the subject's attentional state and to condition the intended speech speed. Two linear mixed models were performed to predict the duration of the produced phonemes and how they were affected by the previously mentioned factors. Results indicate that attention is an influencing factor in both consonants and vowels, with higher levels of attention resulting in longer production times. For the studied consonant (/p/), it was observed that its duration is longer when it is coarticulated with a rounded vowel than when with an unrounded vowel. For vowels, no difference was found in the duration between /a/ and /u/, nevertheless, both modify their duration according to speech rate, with longer times for slower rates.

Keywords: speech production, phoneme variability.

1. 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.

2. 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. A visualisation of the experimental design can be seen in Fig 1.

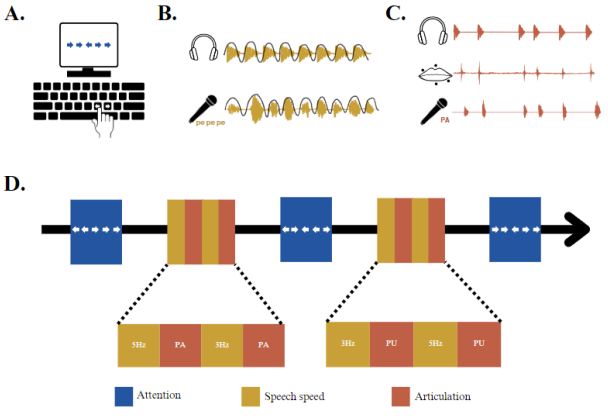


Figure 1. Experimental design. ***A.*** *Flanker task. Participant must indicate the direction of the arrow in the centre (in this case, it is pointing to the left (⇦)).* ***B.*** *Speech speed task. When presented with a rhythmic train of tones at the fast (5 Hz) or slow (3 Hz) condition, participants must whisper the syllable*

/pe/ in synchrony. ***C.*** *Coarticulation task. Lip activity and sound produced by the participants were recorded when they pronounced /pa/ or /pu/ upon getting an audio cue.* ***D.*** *The order of the conditions presented in the speech speed and coarticulation tasks changed randomly for each participant.*

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 estimated as the voiced time obtained from the acoustic signal (i.e. speech offset minus speech onset).

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. No interaction between factors were tested. 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).

3. 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 ms/att.level, $p<0.001$; see Fig. 2A) and shorter duration times when the consequent vowel is /ɑ/ (an unrounded vowel) rather than when the consequent vowel is /u/ (a rounded vowel) ($mean_u$=192 ms, $mean_ɑ$=176 ms, $p<0.001$; see Fig. 2B).

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 ms/% ; $p=0.0016$; see Fig. 2C). 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. 2D).

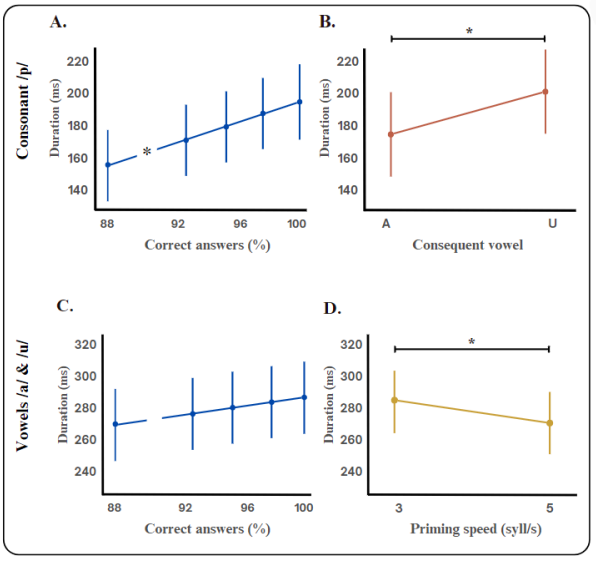


Figure 2. Linear mixed effect models results.. ***A&B.*** *Predicted /p/ duration as a function of attention level and consequent vowel, respectively.* ***C&D.*** *Predicted vowels duration (/a/ and /u/) as a function of attention level and priming speed, respectively. Dots: model predicted group means. Bars: 95% confidence interval. * $p < 0.002$*

4. Discussion and conclusion

We observed that attentional status significantly modulated duration across all the analysed phonemes, with higher attention levels resulting in longer production times. It has been proposed that when attention level is high the articulation of every phoneme is carefully done producing a higher quality speech, which in turn requires longer production times, Dromey and Shim (2008). This can explain 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/, as in a /p/ followed by an /u/, the lips don't only occlude but also round, resulting in longer times than when the incoming vowel is not rounded (as is the /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. This 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).

The presented work expanded our knowledge about the factors that influence speech production. More precisely, we show that: phoneme's articulation time (vowels, as well as consonants) increases with the level of attention; vowel articulation time presents no variability due to their identity, but modifies their duration according to the intended speech rate; and bilabial consonant articulation time is longer when followed by rounded vowels but invariant in respect to the intended speech rate.

5. 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/10.2307/52207

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

Abstract

In fast speech, speech sounds are produced shorter. However, according to previous studies, i) vowels are more compressible, and reduce more than consonants. In languages that show phonemic vowel length contrast, like Japanese, and Hungarian, vowels are also expected to vary in the extent of reduction as a function of their phonological length: in fast speech, ii) long vowels are expected to reduce more than short vowels, while iii) the vowel length contrast (as expressed in duration ratio) does not neutralize completely, as shown for Japanese. In this study, we analyzed consonant and vowel durations produced by 15 Hungarian speakers at comfortable and fast speech rates and tested these three hypotheses.

We found that in fast speech in Hungarian, i) vowels reduced more in their duration than consonants; ii) long vowels reduced more than short vowels; and iii) duration differences of long and short vowels reduced, but duration ratio of the relevant pairs decreased only in the high front pair, while it did not reach complete neutralization in any of the pairs, meaning that the phonologically relevant contrast was maintained across speech rates.

Keywords*: speech production, vowel length, segmental duration, speech rate.*

1. 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), which feature an obstruction in the mouth and can have a complex inner structure. To this claim, we found data from several languages obtained mostly in experiments involving a small number of speakers (Kozhevnikov & Chistovich 1965: 2 Russian speakers; Wood 1973: 2 Swedish speaker, and 1 speaker of British English, Chinese, Polish, German, Egyptian Arabic). But recently, a larger scale corpus study also confirmed it, where segmental durations measured in 20-40 speakers of 8 genetically unrelated languages were obtained in varying conditions using purely automated methods (forced alignment) (Lo & Sóskuthy 2023). With respect to Hungarian, we find no systematic and/or replicable analyses, but there is some evidence supporting the assumption that in fast speech, consonants reduce more than vowels, that is, that consonants are more resistant to speech rate effects than vowels (Magdics 1969).

Differences in segmental reduction in fast speech are also expected according to phonemic length of the segments in languages that exhibit phonological length contrast. Previous data, however, are not entirely conclusive. Japanese and Korean are traditionally described as having phonemic vowel length that is expressed primarily by duration. In Japanese, a study testing speech rate effects found that long vowels were affected more by speech rate than short vowels, that is, they reduced or lengthened more in fast and slow speech, respectively (Hirata 2004). As a result, duration differences of long and short vowel pairs reduced, but duration ratios were maintained in fast speech, reflecting that the vowel length contrast was preserved despite overall reduction tendencies observed in fast speech rates. In Korean, however, short and long vowels were found to be affected similarly across speech rates, that is, they stretched, and reduced to a similar extent in slow, and fast speech, respectively (Magen & Blumstein 1993).

In Hungarian, vowel length is also phonologically distinctive, similar to Japanese and Korean. Phonetic implementation of the contrast is, however, more complex. Traditionally, it is assumed that phonological vowel length is expressed using durational and spectral cues 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/), mainly durational differences can be found between the member of the pairs (Gósy 2004). Previous studies provided some evidence that long vowels are affected more by speech rate than short vowels also in Hungarian (i.e., long vowels reduced to a higher degree than short vowels) (Magdics 1969; Mády 2008), but a more extensive exploration of the realization of the length contrast in fast speech is still warranted.

In accordance with the above, in the present study, we aimed to further advance our knowledge on how segmental reduction and the phonological vowel length contrast is realized in fast speech in Hungarian. We investigated the following three questions: i) are consonants more resistant to speech rate effects and do vowels reduce more in fast speech than consonants? ii) Are long vowels affected more by speech rate than short vowels? iii) Is the difference and the ratio of long and short vowels' duration maintained across different speech rates, in other words, is the phonological length opposition maintained in fast speech?

2. Methods

To answer our research questions, we carried out durational analysis of speech sounds produced in real words. It is important to emphasize in advance, that the experimental material we used in this study was primarily designed for the purposes of controlled, and balanced comparison of long and short vowels. Further, the Hungarian material was also designed to match a corpus of German speech, as our future goal is also to analyze and compare tendencies in these two languages. We opted to use real words which do not constitute minimal pairs, as we expected that this way, the used linguistic material does not facilitate exaggeration of the contrastive

features of segments differentiating minimal pairs, that is, vowel length. This way, we aimed to eliminate some possible factors that would block segmental reduction in fast speech and hoped to get a picture closer to "natural" speech production.

In this study, 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/. To control for place of articulation effects (and create a material comparable in Hungarian and German in a further step), in the onset of the one syllable words we placed laryngeal or alveolar consonants: /h z s t r/. For the same reasons, in the coda, velar and alveolar consonants were positioned: /z t d k n r/.

To control for intonation effects (and again to get comparable data to that we plan to obtain in German in another study), 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: i) at comfortable speech rate ("normal" speech), and ii) at 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 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 accelerating 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. First, we quantified speech rate as word length, and tested if words in fast speech were produced at a higher speech rate than words in normal speech. Then, we analyzed and compared the duration of (long and short) 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 (lme4, Bates et al. 2015; lmerTest, Kuznetsova et al. 2017) followed by post hoc tests (emmeans, Lenth 2021) in R (R Core Team 2018).

3. Results

On average, all speakers were able to speed up their speech considerably. Word durations in fast speech (199.07±64.69 ms) were half of that found in normal speech (396.76±101.98 ms). We also found less variability in fast speech which reflects that speech production at higher speck rates is more demanding for speakers and puts stronger constraints on segment production.

On segmental durations (consonants and vowels pooled), we found a SPEECH RATE*SEGMENT TYPE interaction effect ($F(1, 6390) = 103.69$; $p < .001$), as in normal speech, vowels were inherently longer and they also reduced more in fast speech than consonants (Fig. 1).

We also checked if onset and offset consonants behaved differently (Fig 2). Statistical analysis revealed a significant SPEECH RATE*SYLLABLE POSITION interaction effect ($F(2, 6405) = 146.86$; $p < .001$), and pairwise comparisons showed that while onset and offset consonants had a similar average duration in normal speech, offset consonants reduced more in fast speech than onset consonants. Altogether, we had 3229 obstruents (types /h z s t d k/), 1051 sonorants (types /n r/), and 2155 vowels. Comparison of this unbalanced material showed that in general, sonorant consonants were the shortest in our dataset, but on average they reduced as much as obstruents and less than vowels (MANNER*SPEECH RATE interaction; $F(2, 6359) = 59.12$, $p < .001$) (Fig 3.).

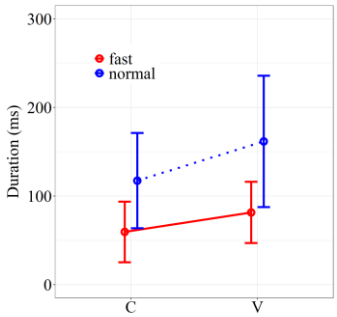


Figure 1: *Vowel and consonant durations in normal and fast speech.*

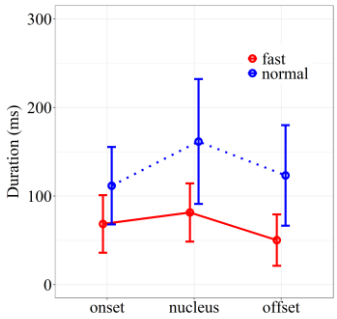


Figure 2: *Vowel and consonant durations as a function of syllable position in normal and fast speech.*

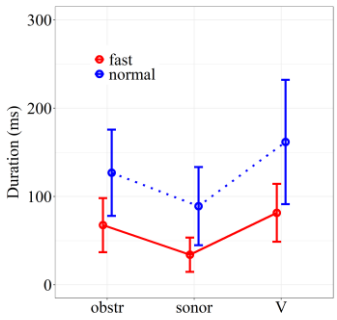


Figure 3: *Vowel and consonant durations as a function of manner of articulation in normal and fast speech.*

On vowel durations, we found a LENGTH*SPEECH RATE interaction effect $F(1, 2095) = 463.34$; $p < .001$), since in normal speech, phonologically long vowels were realized with longer duration (204±57 ms) than short vowels (119±51 ms), while in fast speech, they reduced more (dur_{long} = 95±33 ms; dur_{short} = 69±29 ms; $diff_{long}$= 109 ms; $diff_{short}$ = 50 ms) (Fig. 4.). A larger model, including vowel height as a fixed factor (Fig. 5) also revealed that low vowels, which were inherently longer than high vowels, were also reduced more in

fast speech (LENGTH*HEIGHT*SPEECH RATE interaction $F(1, 2096) = 4.82$; $p < .05$).

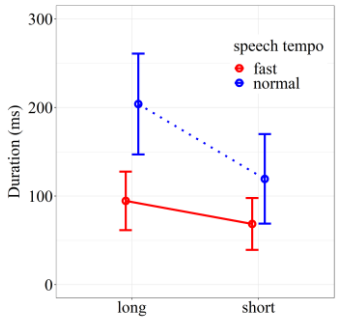


Figure 4: *Vowel durations as a function of phonological vowel length.*

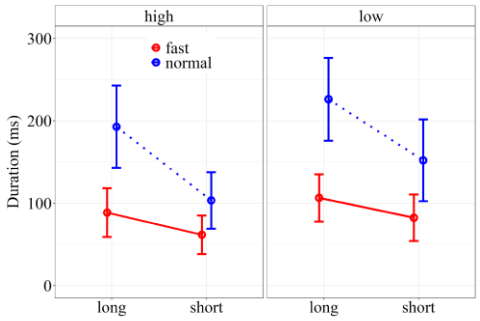


Figure 5: *Vowel durations as a function of phonological vowel length and vowel height.*

Vowel length contrast was analyzed as a function of vowel contrast type, where we had /u/-pairs: /u uː/, /i/-pairs: /i iː/, and /a/-pairs: /ɒ aː/ (Fig. 6 & 7).

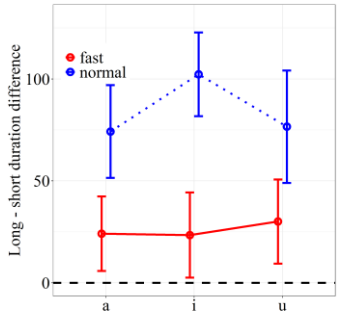


Figure 6: *Differences of long and short vowels' duration as a function of vowel contrast type.*

In normal speech, duration differences of long and short vowels were 74 ms for /a/-pairs, 102 ms for /i/-pairs, and 76 ms for /u/-pairs, on average, that is, front high /i/-pairs showed the greatest duration difference. This was due to the fact that short /i/s were, in this condition, extremely short. In fast speech, these differences were all reduced significantly (statistics showed a VOWEL QUALITY*SPEECH RATE interaction; $F(2, 45) = 17.76$; $p < .01$; but pairwise comparisons confirmed all differences to be significantly smaller in fast speech than in normal speech) (Fig. 6.) (/a/-pairs = 24 ms, /i/-pairs = 24 ms, /u/-pairs = 30 ms). In spite of this extensive durational reduction, differences between corresponding pairs did not reach zero in any of the cases (horizontal dashed black line on Fig. 6.), which would have reflected complete neutralization of the phonological contrast.

Lastly, we turn to the ratio of long and short vowels' duration (Fig. 7.). We found that in normal speech, duration ratio was the highest in the /i/-pair (2.27), followed by the /u/-pair (1.64), and the /a/-pair (1.30). According to statistical analyses, these ratios decreased only in the /i/-pair (VOWEL QUALITY*SPEECH RATE interaction: $F(2, 75) = 16.86$; $p < .01$), where we found the most extreme differentiation of vowels in normal speech, and this differentiation decreased in a way that it became similar to that found in the other two pairs (/i/-pair = 1.5; /u/-pair = 1.46; /a/pair = 1.30). Here, again, we can conclude that none contrast reduced so that it reached complete neutralization (that is, one, numerically; see horizontal dashed black line on Fig. 7.).

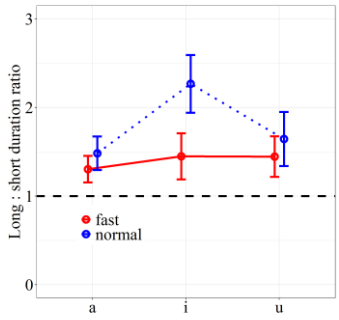


Figure 7: *Ratio of long and short vowels' duration a function of vowel contrast type.*

4. Discussion and conclusions

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 in fast speech, but neither differences, nor the ratio of long and short vowels revealed complete neutralization of the vowel length contrast in fast speech.

These results are in line with expectations, as they support the idea of vowels being more compressible than consonants. But it is also important to emphasize that these findings replicated those documented in previous literature despite the fact that all of these studies used highly varied methodologies.

With respect to phonological length, our data mirrored those of Hirata (2004) for Japanese, showing that phonologically longer segments may be reduced more than shorter segments. On the basis of present results, we may extend this claim and say that physically longer segments may be reduced more than shorter segments. In connection to this, present data also clearly indicated that there is a (probably mechanically motivated) limit to segments compression. This finding may also explain, at least in part, why longer segments may be subjected to more extensive durational reduction in fast speech.

Despite the fact, that in Hungarian spectral and durational cues are used in combination to index phonological vowel length, the length contrast was maintained in duration to some extent in fast speech in basically all cases (irrespective of vowel height) in our Hungarian corpus, similarly to that observed in Japanese where the main cue of vowel length is duration (Hirata 2004). It seems to go against expectations that

we found /i/-pairs to be distinguished the most by durational cues (and not /a/-pairs), as it was documented previously that the vowel contrast in Hungarian low vowels is accompanied by a greater duration difference, and duration ratio, than in high vowels (Deme *et al*. 2019). But this result originates in the fact that /i/ in normal speech was produced extremely short, and this extremity leveled out in fast speech.

As a next step, we plan to analyze spectral differentiation of the same vowel pairs in the two speech rate conditions to test if quality distinction of long and short vowel pairs is also maintained is fast speech. We also analyze data we obtained in a comparable real-word corpus in German speaking females and compare them to the present results. This way, we can test if the same tendencies emerge with respect to consonant and vowel durations, and the vowel length contrast in a language that is typologically not related to Hungarian, and in which vowel length contrast is expressed using similar means, but in a different combination.

The present findings 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.

5. 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.).

6. References

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software,* 67, 1-48.

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

Deme, A., Kohári, A., Reichel, U. D., Szalontai, Á. & Mády, K. (2019). A magánhangzós hosszúsági fonológiai kontraszt a dajkanyelvben a csecsemő életkorának függvényében. [Vowel length contrast in motherese and adult directed speech as a function of the infants age] *Beszédkutatás* 27(1), 221-242.

Gósy, M. (2004). *Fonetika, a beszéd tudománya*. [Phonetics, the science of speech.] Budapest: Osiris.

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*. Washington: Joint Publications Research Service.

Kuznetsova, A. & Brockhoff, P. B., Christensen, R. & Haubo B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26.

Lenth, R. V. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means.* R package. version 1.7.0. https://CRAN.R-project.org/package=emmeans

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.

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

Abstract

Sibilants are characterized by the production of turbulent airflow, which involves both a narrow constriction in the vocal tract and a certain volume velocity of the airflow. Despite both conditions being necessary for sibilant production, studies of constriction degree predominate in the literature. Using acoustic and articulatory data, we show that in certain sequences Standard Mandarin apical vowels exhibit minimal lingual adjustment compared to adjacent sibilants, while also exhibiting a considerable drop in frication noise. The same result was found for the vowel /i/. We hypothesize that the change in frication noise could be due to a number of different non-lingual factors and discuss the potential implications for models of sibilants.
Index Terms: sibilants, apical vowels, Standard Mandarin, ultrasound

1. Introduction

Sibilants are sounds characterized by the production of audible turbulent airflow (Stevens 1998). 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 (Shadle 1990; Catford et al. 1977). These aerodynamic principles suggest that in connected speech the production of frication noise rests on a certain balance being struck between these two factors, e.g. a larger constriction necessitates greater volume velocity and vice versa (Yoshinaga, Nozaki, and Wada 2019). We investigate this relationship between lingual constrictions and aerodynamics in Standard Mandarin apical vowels using both articulatory and acoustic data.

In Standard Mandarin (SM), there is a three-way place contrast among sibilants, with the language contrasting dental, alveolo-palatal and retroflex sibilants, e.g. /s ɕ ʂ/. One consequence of this three-way place contrast is the co-occurrence restriction on the high front vowel /i/ following dental and retroflex sibilants, e.g. *si *ʂi. In these contexts, in place of the high front vowel, there occurs two apical segments, [ɿ] and [ʅ], which occur only after sibilants they are homorganic with, e.g. [sɿ] and [ʂʅ] (Duanmu 2007). [ɿ] will be referred to as the "dental apical vowel" and [ʅ] as the "retroflex apical vowel" in keeping with the previous literature.

Two key characteristics of the apical vowels are the focus of the current study. First, previous research has shown that both apical vowels are produced with a lingual configuration that closely resembles their onsets (Lee-Kim 2014; Faytak and Lin 2015; Shao and Ridouane 2023), though questions remain on the exact nature of the lingual transition from the onset sibilant to apical vowel. While studies have reported a range of adjustments, it is difficult to rule out if any observed differences between the onset and apical vowel were due to coarticulatory effects from the segment following the apical vowel (Foley 2023). Second, there is some debate on whether the apical vowels have frication noise targets (Lee-Kim 2014; Duanmu 2007; Yu 1999; Shao and Ridouane 2023), with few studies firmly quantifying the rate of turbulent airflow during these segments (Shao and Ridouane 2023). While Lee-Kim (2014) concluded that the segments lack frication and termed them "syllabic approximants", Yu (1999) concluded that they are "syllabic sibilants", with both studies using impressionistic inspection of spectrograms and waveforms as evidence.

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.

2. Methods

2.1. Ultrasound experiment

Seven speakers of SM with no history of speech or hearing disorders took part in the study. Data from two speakers was excluded due to errors in the placement of the ultrasound probe. The five remaining speakers were all aged 18-25 years old; three speakers were from northern provinces (Liaoning, Shandong, Shaanxi) and two were from central/southern provinces of China (Henan, Jiangsu).

Stimuli consisted of disyllabic pseudo-words. The target segments in the first syllable are [ɿ ʅ i u], paired with three different onsets [s ʂ ɕ]. Due to phonotactic restrictions, each apical vowel occurs only after homorganic sibilants, [i] occurs only after [ɕ], and [u] occurs only after [s] and [ʂ]. The second syllable is one of [sa ʂa ɕa]. 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 u] in the first syllable used for comparison. All syllables were produced with a high level tone. Sixteen disyllabic filler items were also presented. Stimuli were presented in blocks of five, randomized so that each target phrase was seen a total of five

times across all blocks. The stimuli were presented as simplified Chinese characters in the following carrier phrase: 我觉得___ 很好[wə^{21} tɕyei^{35} də_xən^{35} xau^{213}] "I think _ is very good".

Ultrasound video and audio were co-recorded in a sound-attenuated booth using the Articulate Assistant Advanced (AAA) software. Ultrasound was recorded using a Telemed MicrUs and two different probes, a Telemed MC10 microconvex probe for speakers SP_06 and SP_07 and a Telemed MC4 microconvex probe for all other speakers. Probes were stabilized with a metallic Articulate Instruments stabilization headset. Audio files were analyzed in Praat and segmented using the Montreal Forced Aligner (McAuliffe et al. 2017) with manual corrections as needed.

2.2. Analysis

Zero-crossing rate (ZCR) was used to measure the time course of frication during target sequences. 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 (Shao 2020; Shao and Ridouane 2023). Generalized Additive Mixed Models (GAMMs) (Wood 2011) were constructed to model the dynamics of z-scored ZCR in target sequences, using mgcv v1.8-40 (Wood 2011). We constructed a single model to model all sequences and reported the estimated differences in separate difference figures. In the model, ZCR of [C_1{ɿ, ʅ, i, u}C_2a] sequences was estimated over time, with factor smooths for speaker. Because ZCR has a left-skewed, long-tailed distriubtion, Tweedie distributions were used in the GAMM models. Results were visualized using tidyverse v1.3.2 (Wickham et al. 2019) and tidymv v3.3.2 [1].

Ultrasound frames recorded during the acoustic duration of the target segments were processed in Articulate Assistant Advanced (AAA). Tongue contours were estimated using speaker templates, hand-corrected as necessary, and exported in polar and Cartesian coordinates. To visualize tongue posture over the duration of the target [C_1{ɿ, ʅ}C_2a] items and the comparable [C_1{i, u}C_2a] items, smoothing-spline ANOVAs (SSANOVAs) were generated in polar coordinates comparing the midpoints of the first homorganic fricative (C_1), apical vowel, second homorganic fricative (C_2), and final [a] (Davidson 2006; Gu 2014). The resulting splines and 95% confidence intervals were visualized using tidyverse v1.3.2 (Wickham et al. 2019). The SSANOVAs serve to confirm whether there are any broad adjustments to tongue posture in the transitions between the apical vowels and their flanking homorganic fricatives, and to compare this adjustment to comparison items containing [i u] flanked by the same fricatives.

Additionally, constriction degree (CD) was calculated in AAA using a fiducial line drawn from the probe origin through the alveolar or postalveolar area depending on the constriction at issue. CD was calculated as the distance between the intersections of the fiducial line with the tongue contour and the palate trace. All values were z-scored across speakers. GAMMs were also fit on CD data to model change over the target sequences. The model design was the same as the ZCR models, but the CD models were fit using a Gaussian distribution. Both the ZCR and CD GAMMs were fit using by-phrase relativized time, calculated using $\mathbf{t}_i^{rel} = \mathbf{t}_i - \min(\mathbf{t})/\max(\mathbf{t}) - \min(\mathbf{t})$, where $\mathbf{t}_i$ is a single timepoint.

[1] https://stefanocoretta.github.io/tidymv

3. Results

3.1. SSANOVAs

The SSANOVA splines in Figure 1 summarize the typical posture for the imaged portion of the tongue at the midpoint of each segment in target [C_1{ɿ, ʅ}C_2a] items, with [ɕi.ʂa] shown for comparison. In the apical vowel targets, 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]. Unexpectedly, the tongue blade is also raised at the midpoint of [i] for all speakers, not appreciably differing from the raising observed for [ɕ]; in fact, the tongue postures of [ɕ] and [i] are essentially the same, except for speakers SP_02 and SP_05 who show somewhat more dorsum raising during [i].

3.2. Zero-crossing rate

Figure 2 shows the results from the time-aligned ZCR (bottom) and CD (top) GAMMs. Constants were added to both sets of values for visualization. Two clear peaks in ZCR corresponding to the sibilants [s ɕ ʂ] are visible in the targets, as well as two valleys corresponding to the nuclei [a u ɿ ʅ]. The ZCR values in V_1 position are consistently much lower compared to the two flanking peaks, suggesting that V_1 has reduced aperiodicity compared to [s ɕ ʂ]. Crucially, while we can see clear differences in ZCR between the four phrases during the two flanking sibilants, the ZCR trajectories all converge to a common minimum near the V_1 midpoint.

The GAMM estimates of difference in ZCR are shown in Figure 3, where shaded red regions show intervals during which this comparative difference is significant. In most cases, the difference in aperiodic noise is significantly different during the two sibilants, i.e. C_1 and C_2, with a gap in this difference during the midpoint of V_1. Interestingly, the former is true even in the comparison when both phrases have the same sibilants (top panel). This suggests that the different nuclei in each phrase have a direct impact on how favorable the conditions are for the production of turbulent airflow, with the dental apical vowel creating a more favorable context. This is likely related to the homorganicity between the onset [s] and apical vowel [ɿ].

3.3. Constriction degree

GAMMs fit on the CD data are shown in Figure 2 (top) for the target phrases. It can be seen clearly that for all of the phrases with homorganic sequences, i.e. those with the apical vowels and [i], a consistent CD is maintained during the first sibilant and V_1, with the constriction being released during the C_2 in anticipation of the following [a]. For the other phrase containing [u] as V_1, a sizeable dip in CD occurs in accord with the V_1 onset, only for a subsequent constriction to be formed for the second sibilant. This is indicative of a slight release in the tongue front constriction during the production of the vowel [u] in this phrase. This slight release in constriction coincides with the drop in frication seen in the ZCR GAMM, while for the other phrases, there is no perceptible change in CD that coincides with the change in aperiodic noise.

The GAMM estimates of difference in CD are shown in Figure 4. In the comparisons in panels 1 and 3 (1 being the top panel), where the phrases with the apical vowels are compared to those containing the phrase with [u], we see a period of significant difference during the drop in CD that occurs during the

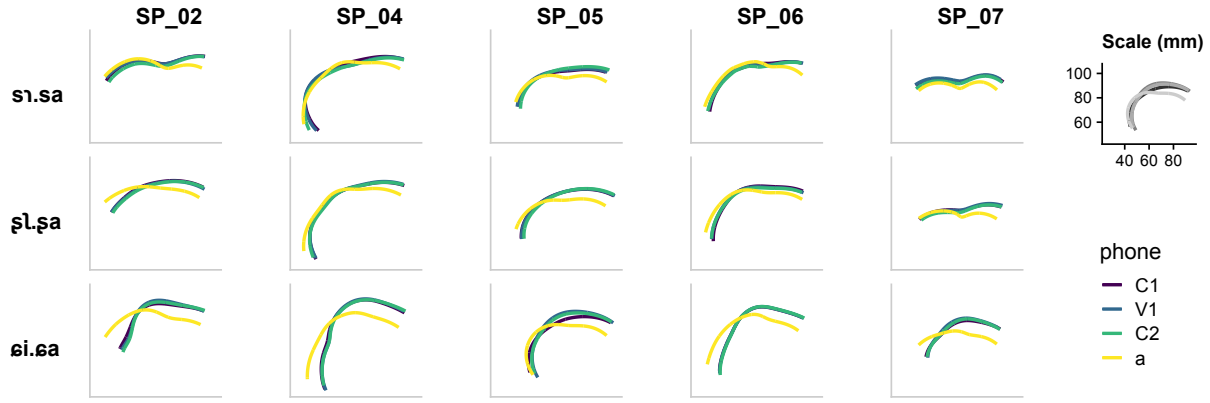


Figure 1: Tongue surface SSANOVA splines for segment midpoints in target [$C_1V_1C_2$a] items. Anterior is right in each figure.

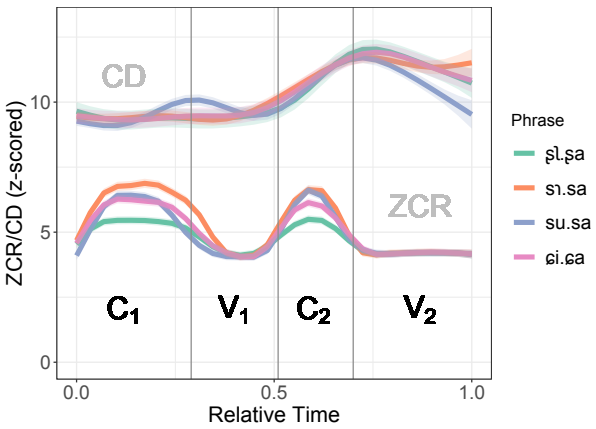


Figure 2: ZCR (bottom) and CD (top) GAMMs for all target sequences. Grey vertical lines indicate phone boundaries.

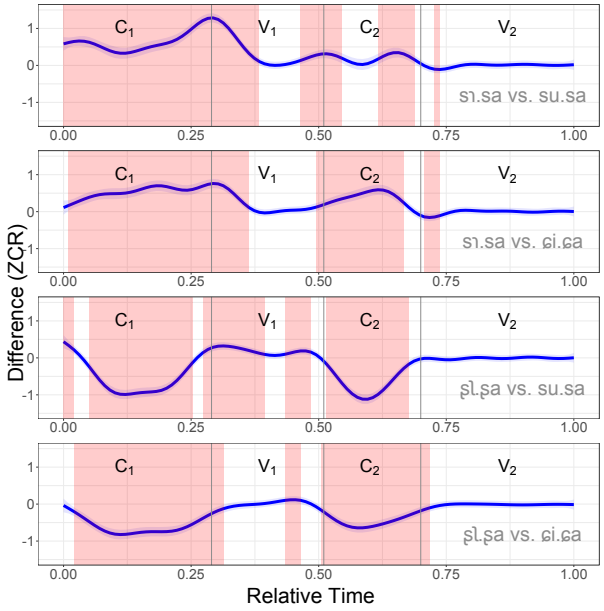


Figure 3: GAMM difference smooths for ZCR. Grey vertical lines indicate phone boundaries. Red shaded regions indicate regions of statistically significant difference.

onset of [u]. Interestingly, in the [sɻ.sa] versus [su.sa] comparison, there is also a period of difference during the formation of the second sibilant, with a more narrow constriction formed during the latter phrase. In the two comparisons between homorganic sequences, i.e. panels 2 and 4, the differences are near zero for the entirety of the duration, indicating that the changes in CD during these phrases follow very similar trajectories.

4. 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 [$C_1V_1C_2$a] items, a considerable drop in frication occurs during apical vowels in V_1 position, following nearly the same trajectory as the other vowels examined. Second, no change in CD occurs during the apical vowels in the target sequences, as confirmed by examination of tongue posture at segment midpoints and kinematic analysis over the whole duration of the items. 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 an increase in channel size.

During the target sequences, speakers may turn towards some non-lingual adjustment to suppress frication during V_1 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 (Iskarous, Shadle, and Proctor 2011; Recasens, Pallarès, and Fontdevila 1997). One potential hypothesis is that speakers are directly manipulating the rate of airflow in the vocal tract during the apical vowels in V_1 positions. This would indicate the presence of airflow velocity targets separate from constriction degree targets, suggesting that gestural approaches to phonology that only incorporate constriction degree targets are overly simplistic (Iskarous, Shadle, and Proctor 2011; Browman and Goldstein 1989).

Alternatively, one could argue that the drop in frication dur-

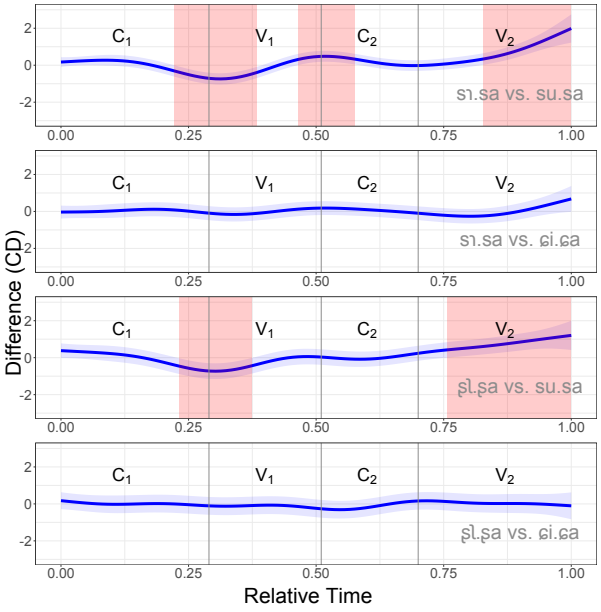


Figure 4: GAMM difference smooths for CD. Grey vertical lines indicate phone boundaries. Red shaded regions indicate regions of statistically significant difference.

ing the apical vowels and [i] is merely due to the onset of voicing. The antagonistic relationship between voicing and frication could potentially lead to a drop in the rate of turbulent airflow during the apical vowels (Ohala and Solé 2010). However, to maintain the position that the apical vowels have frication noise targets, this predicts that the overall rate of frication during the apical vowels should be *higher* than that of other vowels, as reported for the Jixi apical vowel (Shao and Ridouane 2023). The current results show no significant difference in the trajectory of frication noise during the apical vowel sequences compared to that of the other vowels. Incorporating the voiced fricative [z̩] before the apical vowel [ɻ] into the stimuli would allow for testing this hypothesis (e.g. [z̩ɻ.z̩u]). If the trajectory of frication during these sequences does not differ from those observed here, that would suggest other mechanisms are at play here.

In conclusion, this study looked at the trajectory of frication and CD during sequences containing SM apical vowels and sibilants they are homorganic with in comparison to other vowels in the same sequences. The results showed little to no adjustment in CD during the apical vowels in these sequences, with a considerable drop in frication during this same period. Given that turbulency requires both a certain channel size and airflow velocity, we hypothesize that some adjustment is suppressing the rate of airflow during these sequences. This leaves open the possiblity that speakers are directly manipulating the rate of airflow, though other adjustments are possible. Further investigation is needed in these regards.

5. Acknowledgments

This work was supported by NIH grant T32 DC009975 (Foley).

6. References

Browman, Catherine P and Louis Goldstein (1989). "Articulatory gestures as phonological units". In: *Phonology* 6.2, pp. 201–251.

Catford, John Cunnison et al. (1977). *Fundamental problems in phonetics*. Midland Books.

Davidson, Lisa (2006). "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance". In: *The Journal of the Acoustical Society of America* 120.1, pp. 407–415.

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

Faytak, Matthew and Susan Lin (2015). "Articulatory variability and fricative noise in apical vowels." In: *ICPhS*.

Foley, Sean (2023). "The coarticulatory behavior of Standard Mandarin apical vowels". In: *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023)*.

Gu, Chong (2014). "Smoothing Spline ANOVA Models: R Package gss". In: *Journal of Statistical Software* 58.5, pp. 1–25. URL: https://www.jstatsoft.org/v58/i05/.

Iskarous, Khalil, Christine H Shadle, and Michael I Proctor (2011). "Articulatory–acoustic kinematics: The production of American English/s". In: *The Journal of the Acoustical Society of America* 129.2, pp. 944–954.

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.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." In: *Interspeech*. Vol. 2017, pp. 498–502.

Ohala, John J and Maria-Josep Solé (2010). "Turbulence and phonology". In: *Turbulent sounds: An interdisciplinary guide*, pp. 37–97.

Recasens, Daniel, Maria Dolors Pallarès, and Jordi Fontdevila (1997). "A model of lingual coarticulation based on articulatory constraints". In: *The Journal of the Acoustical Society of America* 102.1, pp. 544–561.

Shadle, Christine H (1990). "Articulatory-acoustic relationships in fricative consonants". In: *Speech production and speech modelling* 55, pp. 187–209.

Shao, Bowei (2020). "The apical vowel in Jixi-Hui Chinese: phonology and phonetics". PhD thesis. Université Sorbonne Nouvelle.

Shao, Bowei and Rachid Ridouane (2023). "On the nature of apical vowel in Jixi-Hui Chinese: Acoustic and articulatory data". In: *Journal of the International Phonetic Association*, pp. 1–26.

Stevens, Kenneth N (1998). *Acoustic phonetics*. MIT press.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.

Wood, S. N. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". In: *Journal of the Royal Statistical Society (B)* 73.1, pp. 3–36.

Yoshinaga, Tsukasa, Kazunori Nozaki, and Shigeo Wada (2019). "A simplified vocal tract model for articulation of [s]: The effect of tongue tip elevation on [s]". In: *PloS one* 14.10, e0223382.

Yu, Alan CL (1999). "Aerodynamic constraints on sound change: The case of syllabic sibilants". In: *The Journal of the Acoustical Society of America* 105.2, pp. 1096–1097.

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 de 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; Research School for Behavioral and Cognitive Neurosciences (BCN), 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

Abstract

Speech-Music Therapy for Aphasia (SMTA) is applied in the treatment of childhood apraxia of speech. A single-subject design study into the effect of SMTA was conducted, showing a striking improvement in the realization of initial consonant clusters after treatment.
Musical scores of the target items with initial consonant clusters were analysed to determine the musical manipulations that were used.
In all target items with initial consonant clusters an anacrusis was used. This musical manipulation may have provided an auditory rhythmic cue for the phasing relationships of speech gestures in consonant clusters.
Further research is needed to determine effective musical manipulations that are used in SMTA

Keywords: speech production, treatment, music, childhood apraxia of speech.

1. Introduction

Speech-Music Therapy for Aphasia (SMTA; De Bruijn et al., 2005) is applied in the treatment of childhood motor speech disorders (van Tellingen et al., 2023). The use of music in the treatment of motor speech disorders is supported by theories on similarities between speech and music. In SMTA, musical manipulations, such as change in tempo or rhythm and speech therapy manipulations, such as visual cues, are systematically applied to improve speech production. Results of a single-subject design study into the effect of SMTA in the treatment of childhood apraxia of speech showed improved realisation of initial consonant clusters after treatment with SMTA (van Tellingen et al., 2023). In the current study we aim to explain the improvement in the realization of consonant clusters by exploring a musical manipulation that was used during treatment in the single-subject design study.

1.1. Childhood motor speech disorders

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 that more severe cases of SSD often present with a combination of problems in linguistic and motor speech processes.
The severe cases of SSD that include problems in motor speech processes are classified as motor speech disorders (MSD). The severity of MSD is reflected in the persistency of the disorder and potential lifelong effects on social, academic and vocational aspects of life (Shriberg et al., 2010).
Childhood apraxia of speech is a specific subtype of MSD in which an impairment at the level of speech motor planning and programming causes prosodic and speech sound production errors (American Speech-Language-Hearing Association, 2007; Shriberg et al., 2010). CAS is defined in three core features: inconsistency, inappropriate prosody and disrupted coarticulation (American Speech-Language-Hearing Association, 2007; Terband et al., 2019).

1.2. Treatment of CAS

Treatment methods for CAS can be roughly divided into two categories, although most treatments include manipulations from both categories. Examples of more articulatory-kinematic approaches are Dynamic Temporal and Tactile Cueing (Strand, 2020) and Rapid Syllable Transition Training (Ballard et al., 2010; McCabe et al., 2014). These methods focus on facilitating more accurate movement through manipulations such as visual and tactile cues and providing specific feedback about the movement (Ballard et al., 2010; Strand, 2020). Examples of more rate/rhythm control type approaches are Melodic Intonation Therapy (Albert et al., 1973; Helfrich-Miller, 1984) and Speech-Music Therapy for Aphasia (SMTA; De Bruijn et al., 2005; Hurkmans et al., 2015; van Tellingen et al., 2023). These methods apply manipulations that focus on rhythm and fluency, such as reduced speech rate and rhythmical cueing (Albert et al., 1973; Hurkmans et al., 2015). Both MIT and SMTA were originally developed in the treatment of adults with aphasia and/or apraxia of speech and were later applied in the treatment of children with CAS.
The current study focusses on the musical manipulations in SMTA, therefore, this method and the rationale for the use of music in the treatment of MSD will be described in more detail in the next sections.

1.3. Speech-Music Therapy for Aphasia

SMTA is a combination of 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 direct imitation, and ending with response to a question (see van Tellingen et al. (2023) for a detailed description).

1.4. Speech and music

The use of music in the treatment of childhood MSD 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 both sound envelope processing in perception of music and speech, and synchronization and entrainment to a pulse in rhythm production in music and speech.

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). By adding music to the treatment, these parameters can be used to enhance differences in pitch, duration and intensity, and improve prosody.

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 1**), showing phasing and duration for gestures in the vocal tract variables that are modulated.

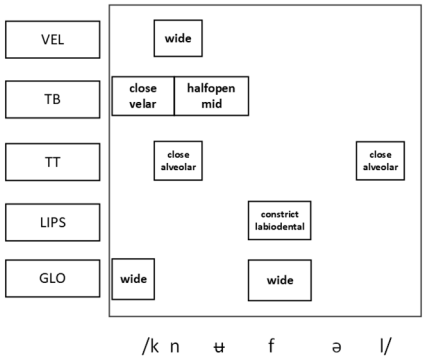


Figure 1: *Gestural score for target item /knʉ-fəl/ (plush toy). VEL=velum, TB=tongue body, TT=tongue tip, GLO=glottis.*

In a similar manner, the notes in a melody are organised in time. **Figure 2** shows a visual representation of this phasing and duration of musical notes for multiple instruments (vocals and guitar) in a musical score.

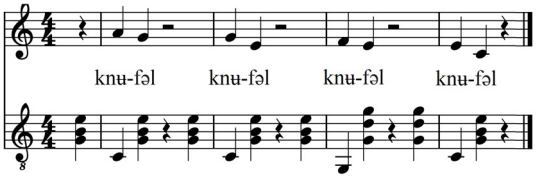


Figure 2: *Musical score for vocals and guitar for the SMTA melody for /knʉ-fəl/.*

1.5. SMTA in the treatment of CAS

The effect of SMTA in the treatment of CAS was evaluated in a single-subject design study. The participant in this study was a 5-year old boy with CAS. He presented with inconsistent speech in both spontaneous speech and repetition tasks. Another speech feature was increasing problems with increasing complexity, which was specifically apparent in the realisation of consonant clusters across speech tasks. Further features included syllable segmentation, groping, consonant deletion and substitution and initial consonant elongation.

Treatment consisted of two 30-minute sessions of SMTA per week for ten weeks. Target items were drawn up to be fitting for both the speech targets, including consonant clusters, and to be functionally relevant for the boy.

A comparison of pre-test, post-test and follow-up measures showed that the boy's intelligibility in daily live improved after treatment. Treatment effects generalized to untrained speech tasks with improvement in the realisation of consonants, vowels and consonant clusters in spontaneous speech, picture naming, non-word repetition, and diadochokinesis. The boy gained the syllable structure CCVC after treatment which was maintained at follow-up (van Tellingen et al., 2023).

In the single-subject design study on SMTA by Van Tellingen et al. (2023) we found a striking improvement in the realisation of clusters in picture naming, non-word imitation, spontaneous speech and diadochokinesis after treatment. In the present study we explored the musical manipulations that were used to support the production of initial consonant clusters to explain this treatment effect.

2. Methods

Musical scores for the treated items in the single-subject design were collected from the performing music therapist. The melodies for items with initial consonant clusters were selected for further analysis. All treated items with initial clusters are presented in **Table 1**. The melodies for these items were written out in musical scores by the music therapist.

Table 1: *Trained items with initial consonant clusters*

Item (Dutch)	Phon. transcription	English translation
Straks	strɑks	*Later*
Groep	xrup	*Group*
Drie	dri	*Three*
Knutselen	knʉtsələn	*arts and crafts*
Kleien	klɛijən	*play dough*
Strijkkralen	strɛikralən	*ironing beads*
Klein	klɛin	*Little*
Greet	xret	*teachers name*
Drenthe	drɛntə	*name of Dutch province*
Skaten	sketən	*rollerskating*
Schaatsen	sxatsən	*iceskating*
Springen	sprɪŋən	*to jump*
Trampoline	trɑmpolinə	*trampoline*
Knuffel	knʉfəl	*stuffed animal*

The first author and the music therapist wrote out the musical manipulations that were applied in musical scores, leading to two or multiple musical scores per trained item. The musical scores with musical manipulations were analysed for similarities and differences.

3. Results

The main musical manipulation in melodies for items including initial consonant clusters was an anacrusis (pickup; a note that precedes the first beat of a measure). An example is presented in **Figure 3,** showing the melody and guitar accompaniment for the treated item /knʉ-fəl/ (plush toy in Dutch).

The duration of the anacrusis was manipulated in several items to vary the timing of sequential realisation of the consonants in the cluster and thereby decreasing or increasing difficulty. This led to lengthening of one of the consonants (in the example the /n/) or insertion of a schwa between consonants in the cluster (in the example this would lead to /kə-nʉ-fəl/). This variation is represented in the musical scores in **Figure 4.** When both consonants were produced correctly, duration of lengthening or schwa was reduced to achieve normal timing relations in clusters as represented in **Figure 3**.

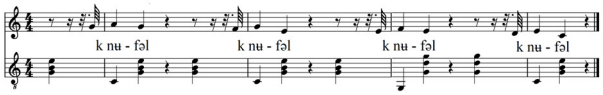


Figure 3: *Musical score for vocals and guitar for /knʉ-fəl/ including an anacrusis*

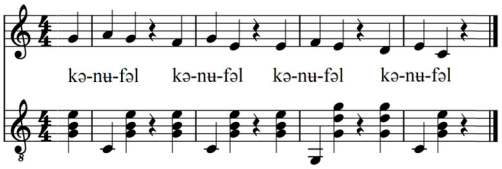


Figure 4: *Musical score for vocals and guitar /kə-nʉ-fəl/ with an anacrusis.*

In another example, /sχa:tsə/ (ice-skating), the anacrusis was used to elongate the initial /s/, which provided time to make the movement towards the next consonant /χ/ (velar fricative). In addition to the anacrusis, rate was also used to provide more time for the realization of both consonants in the cluster by slowing down (rallentando) to elongate the /s/ in the anacrusis and then return to normal tempo (accelerando) for the remainder of the word. With increased ability to produce both consonants correctly, length and tempo of the anacrusis were adjusted to reach normal speech rate.

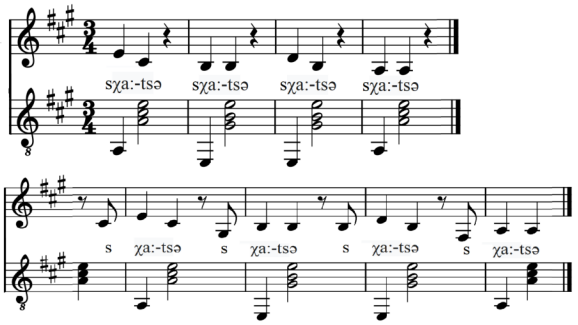


Figure 5: *Basic melody and melody with anacrusis for vocals and guitar for the item /sχa:tsə/ (iceskating).*

A third example is a word with an initial cluster consisting of three consonants /sprIŋə/ (to jump). In this case the boy reduced this cluster by deleting /r/. Therefore, /s/ and /p/ were combined in the anacrusis and the first beat was elongated through slowing down to allow for elongation of /ɪ/. Again, with increased ability, the anacrusis and tempo were adjusted to support combining of all three consonants in the cluster.

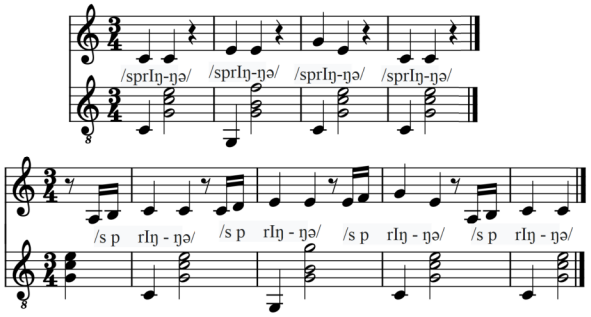


Figure 6: *Basic melody and melody with anacrusis for vocals and guitar for the item /sprIŋə / (to jump).*

4. Discussion and conclusion

In this study we explored the musical manipulations in SMTA that may have contributed to improved realisation of initial consonant clusters in a boy with CAS. Two specific manipulations were used, the addition of an anacrusis for one element of the cluster and slowing down (rallentando) on the consonant cluster.

The use of the anacrusis and rallentando both appear to be used to influence rate, overall (rallentando) and within the consonant cluster (anacrusis). However, the anacrusis also ads an event in the musical phrase, which represents a speech event, the first consonant in the cluster. In **Figure 7** this representation is visualised by aligning the gestural score and musical score including the anacrusis for the item /knʉ-fəl/. The overlap in the phasing and duration of elements involved in both activities may have contributed to the facilitatory effect of the musical manipulations in this case. The anacrusis potentially provided an auditory rhythmic cue for the phasing relationships of speech gestures in consonant clusters.

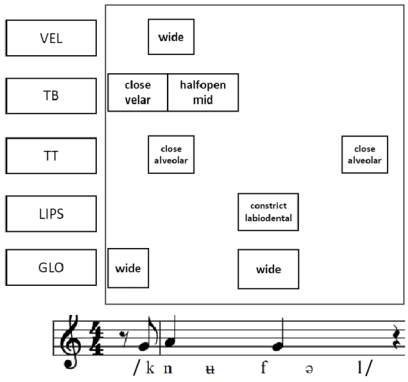


Figure 7: *Aligned gestural and musical scores for /knʉ-fəl/.*

The same musical manipulation, anacrusis, was used in all consonant clusters that were trained. However, the effect of this manipulation on the timing of the realisation of each consonant in the cluster differed, due to characteristics of the trained consonants. For instance, if the first consonant is a stop, as in /knʉ-fəl/, elongation is not possible. If more time is needed, this will lead to /ə/-insertion. In a cluster that starts with a fricative, like /sχa:tsən/, elongation is easier and /ə/-insertion will most likely be avoided.

Another factor is voicing. Both /k/ and /s/ are voiceless, which means they can't be sung on a specific pitch. However in the melodies, the anacrusis always differs in pitch from the beat. Musically, this is logical as the anacrusis leads up to the beat. When a /ə/ is inserted it becomes clear that this pitch difference contributes to placement of stress on the beat. In this way, in

items where the voiceless first consonant can't be sung, the musical accompaniment still supports prosodic features of the target item.

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. Rhythm in music is more predictable than in speech, making it a better stimulus for either entrainment or tracking (Fiveash et al., 2021). The predictable framework that is set up in the musical support in SMTA, could contribute to the treatment effect by providing a predictable framework for the timing of the speech gestures. 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.

4.1. Limitations

This study has some limitations. The first limitation concerns the size of the study. The melodies that were analysed came from one single-subject design study. Analysis of melodies from more studies with variation in speech targets and musical manipulations is needed to draw conclusions on the effect of specific musical manipulations.

The second limitation of this study is found in the approach. In this exploratory study we used materials that were not controlled. This provided the opportunity to explore a surprising finding from previous research. To further determine the effect of these (and other) musical manipulations, future research should focus on direct comparison of manipulations or applying manipulations in a more controlled way to evaluate their effect.

4.2. Conclusion

An exploration of the musical manipulations in SMTA showed that the use of an anacrusis could explain the treatment effect on the realisation of initial consonant clusters. The anacrusis potentially provided an auditory rhythmic cue for the phasing relationships of speech gestures in consonant clusters. Further research is needed to assess the musical manipulations in larger groups of children with different speech targets and varying musical manipulations.

5. Acknowledgements

We thank Ariska Groen (Music Therapist) for providing the melodies and assisting in the analysis.

6. References

Albert, M. L., Sparks, R. W., & Helm, N. A. (1973). Melodic Intonation Therapy for Aphasia. *Archives of Neurology*, *29*(2), Article 2.

American Speech-Language-Hearing Association. (2007). *Childhood apraxia of speech [Technical report]. Available from www.asha.org/policy.* American Speech-Language-Hearing Association.

Ballard, K. J., Robin, D. A., McCabe, P., & McDonald, J. (2010). A Treatment for Dysprosody in Childhood Apraxia of Speech. *Journal of Speech, Language, and Hearing Research*, *53*(5), Article 5.

Browman, C. P., & Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica*, *49*(3–4), 155–180. https://doi.org/10.1159/000261913

De Bruijn, M., Zielman, T., & Hurkmans, J. J. S. (2005). *Speech-Music Therapy for Aphasia (SMTA)*. Revalidatie Friesland.

Diepeveen, S., Terband, H., van Haaften, L., van de Zande, A. M., Megens-Huigh, C., de Swart, B., & Maassen, B. (2022). Process-Oriented Profiling of Speech Sound Disorders. *Children*, *9*(10), 1502.

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. *Front Hum Neurosci*, *8*(OCT), Article OCT.

Haegens, S. (2020). Entrainment revisited: A commentary on Meyer, Sun, and Martin (2020). *Language, Cognition and Neuroscience*, *35*(9), 1119–1123.

Helfrich-Miller, K. R. (1984). Melodic intonation therapy with developmentally apraxic children. In Seminars in Speech and Language (Vol. 5, No. 02, pp. 119-126). © 1984 by Thieme Medical Publishers, Inc.. *Seminars in Speech and Language*, *5*(2), Article 2.

Hurkmans, J. J. S. (2016). *The treatment of apraxia of speech*. Rijksuniversiteit Groningen.

Hurkmans, Jonkers, R., Bruijn, M. de, Boonstra, A. M., Hartman, P. P., Arendzen, H., & Reinders-Messelink, H. A. (2015). The effectiveness of Speech–Music Therapy for Aphasia (SMTA) in five speakers with Apraxia of Speech and aphasia. *Aphasiology*, *29*(8), Article 8.

Hustad, K. C. (2012). Speech Intelligibility in Children With Speech Disorders. *Perspectives on Language Learning and Education*, *19*(1), Article 1.

McCabe, P., Macdonald-D'Silva, A. G., Rees, L. J. van, Ballard, K. J., & Arciuli, J. (2014). Orthographically sensitive treatment for dysprosody in children with Childhood Apraxia of Speech using ReST intervention. *Developmental Neurorehabilitation*, *17*(2), Article 2.

McLeod, S., & Baker, E. (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. *Hearing Research*, *308*, 98–108.

Shriberg, L. D., Fourakis, M., Hall, S. D., Karlsson, H. B., Lohmeier, H. L., McSweeny, J. L., Potter, N. L., Scheer-Cohen, A. R., Strand, E. A., Tilkens, C. M., & Wilson, D. L. (2010). Extensions to the Speech Disorders Classification System (SDCS). *Clinical Linguistics & Phonetics*, *24*(10), 795–824.

Strand, E. A. (2020). Dynamic Temporal and Tactile Cueing: A Treatment Strategy for Childhood Apraxia of Speech. *American Journal of Speech-Language Pathology*, *29*(1), Article 1. https://doi.org/10.1044/2019_AJSLP-19-0005

Terband, H., Namasivayam, A., Maas, E., van Brenk, F., Mailend M.L., Diepeveen, S., van Lieshout, P., & Maassen, B.. (2019). Assessment of Childhood Apraxia of Speech: A Review/Tutorial of Objective Measurement Techniques. *Journal of Speech, Language, and Hearing Research*, *62*(8S), 2999-3032.

Thaut, M., & Abiru, M. (2010). Rhythmic Auditory Stimulation in Rehabilitation of Movement Disorders: A Review Of Current Research. *Music Perception*, *27*(4), 263–269.

van Tellingen, M., Hurkmans, J., Terband, H., van de Zande, A. M., Maassen, B., & Jonkers, R. (2023). Speech and Music Therapy in the Treatment of Childhood Apraxia of Speech: An Introduction and a Case Study. *Journal of Speech, Language, and Hearing Research*, 1–19.

Acoustic correlates of the nasal vs. plosive quantity contrast in Hungarian

Tilda Neuberger

HUN-REN Hungarian Research Centre for Linguistics, Hungary

neuberger.tilda@nytud.hun-ren.hu

Abstract

This study investigates the phonetic realization of consonant length in Hungarian. It is hypothesized that spectral structure differences between obstruents and sonorants may lead to distinct strategies in expressing quantity contrast. To test this hypothesis, intervocalic nasals (/n ɲ/) and plosives (/t k/) were analyzed in spontaneous speech from 20 monolingual Hungarian-speaking adults. Linear mixed-effects models and decision trees were applied to explore the effect of quantity, consonant type, and their interaction on various acoustic parameters, such as the durations of the target consonants and neighboring vowels, relative durations, and geminate-to-singleton ratio. Our findings indicate that nasals require more robust adjustments compared to plosives in the realization of the consonant length contrast. This study contributes to the understanding of phonetic variation in Hungarian and the distribution of geminates across languages.

Keywords: speech production, consonant length, nasal, stop, Hungarian

1. Introduction

Length serves as a distinctive 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). It is claimed that the primary acoustic correlate of geminates is the increased duration of the closure or constriction. However, findings concerning other potential attributes of length, such as preceding vowel duration, voice onset time, fundamental frequency, or amplitude, are not consistent across languages (Al-Tamimi & Khattab 2018; Lahiri & Hankamer 1988; Ridouane 2010).

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. Until now, investigations into length contrast have concentrated on Hungarian plosive consonants (e.g., Deme *et al.* 2018; Neuberger 2023). No study has yet undertaken a comparison across different types of consonants in this regard.

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.

2. Methods

Intervocalic nasal /n ɲ/ and plosive /t k/ singletons and geminates (N = 427) were collected from the spontaneous speech of 20 monolingual Hungarian-speaking adults (10 males) using the BEA database (Neuberger *et al.* 2014). The number of singleton and geminate consonants was quasi-balanced within each consonantal category. There was an attempt to exclude variation due to phonetic factors. Specifically, words containing target segments were selected to have a syllable count ranging from 2 to 4, while excluding initial and final segments. Regarding geminate types (see Ridouane 2010; Neuberger 2023), only lexical and word-internal assimilated geminates were considered, with concatenated geminates being excluded from the analysis. Surrounding vowels were short /ɒ ɛ o/.

The following acoustic parameters were measured by means of Praat (Boersma & Weenink 2020):

- Absolute duration of the target consonant (C): total duration of nasals and plosives (including closure duration, burst and release phase, i.e., voice onset time in case of voiceless plosives).
- Absolute duration of the preceding (V1) and the following vowel (V2): The segmentation of the vowels was based on their second formants supported by visual analysis display of the spectrograms and oscillograms.
- Relative duration of consonants and vowels (C/V1, C/V2): duration related to preceding and following vowel duration.
- Geminate-to-singleton ratio (G/S): durational ratio calculated by each consonant and by each speaker.

Instead of the raw durations, we used the logarithmic values of the absolute consonant and vowel durations because it is suggested that logarithmic durations are relational invariant acoustic variables that can cope with the durational variations of singleton and geminate consonants in a wide range of speaking rates (Amano *et al.* 2021).

Linear mixed-effects models (lmer and lmerTest packages: Bates *et al.* 2014; Kuznetsova *et al.* 2017) were constructed using R (R Core Team 2018) for each acoustic parameter to investigate the effect of quantity (singleton vs. geminate), consonant type (nasal vs. plosive) and their interaction. The random factor was the speakers (N = 20). The effect of gender contributed no improvement to the models and was thus excluded during model selection. Pairwise comparisons with Tukey method were performed with emmeans (Lenth 2018). F-values and corresponding p-values were computed using the

Satterthwaite method. Plots were made with the ggplot2 package (Wickham 2016).

Additionally, decision trees were employed to identify the most important features in distinguishing the two phonological length categories in nasals and plosives. The models were trained on the following variables: logCdur, logV1dur, logV2dur, C/V1, C/V2. Decision trees were constructed using scikit-learn 1.4.2 in Python (Pedregosa *et al.* 2011).

3. Results

Our 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: $F(2, 426) = 9.836$; $p = 0.002$. According to the Tukey post-hoc analysis, statistically significant differences were observed between nasals and plosives for both singletons and geminates ($p < 0.001$ in both cases). Nasals were produced with significantly shorter durations compared to plosives.

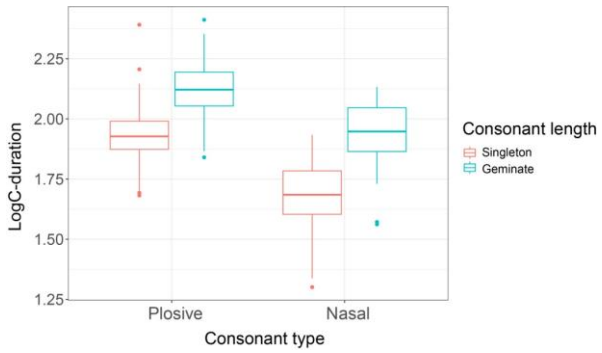


Figure 1: *Consonant duration (log-transformed) as a function of consonant length and consonant type.*

The G/S ratio was significantly higher for nasals compared to plosives, indicating a more distinct contrast in nasals. On average, it was 1.89 (±0.4) for nasals and 1.57 (±0.1) for plosives.

Preceding vowel duration (V1) showed discrepancies between nasal and plosive quantity contrasts (Figure 2). A significant interaction between consonant quantity and consonant type on V1 duration was observed: $F(2, 426) = 25.338$; $p < 0.001$. Vowel duration varied depending on the following consonant type. V1 was longer before nasal geminates compared to nasal singletons. This difference, however, was not observed with plosives, as V1 durations exhibited similar patterns before both singletons and geminates.

The duration of the following vowel (V2) differed significantly between plosives and nasals ($F(2, 426) = 6.431$; $p = 0.011$) but consonant length did not have an effect on this variable (Figure 3). Vowels following nasals were longer then after plosives, on average. To sum up, acoustic results also showed that the duration of the surrounding vowels helps distinguish the two phonological categories, with a greater contribution shown for nasals. The duration of the following vowels showed an opposite trend according to consonant type: for plosives, it was shorter after geminate than after singleton, while for nasals it was the other way round, longer after geminate than after singleton.

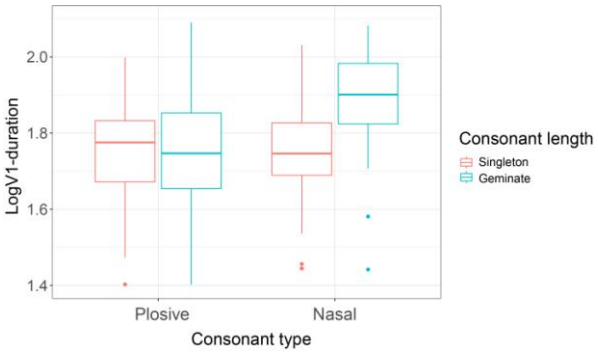


Figure 2: *Preceding vowel duration (log-transformed) as a function of consonant length and consonant type.*

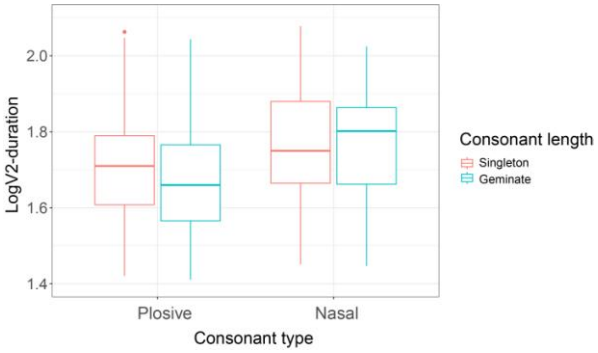


Figure 3: *Following vowel duration (log-transformed) as a function of consonant length and consonant type.*

Considering relative durations, a significant interaction between consonant quantity and consonant type on the ratio of consonant and preceding vowel duration (C/V1) was observed: $F(2, 426) = 13.414$; $p < 0.001$. In terms of this parameter, nasals and plosives differed significantly both for singletons ($p = 0.001$) and geminates ($p < 0.001$). Based on the Tukey post-hoc analysis, a statistically significant difference between singletons and geminates was identified exclusively for plosives ($p < 0.001$), whereas no significant contrast was observed for nasals in this parameter (Figure 4).

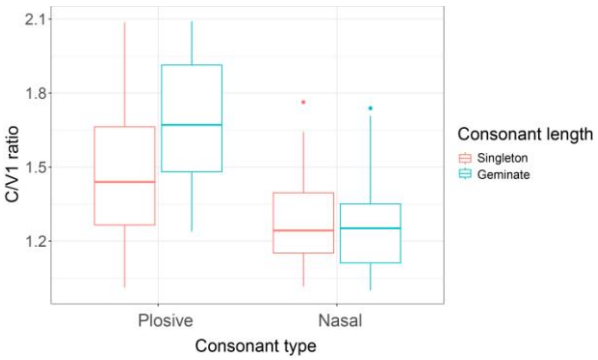


Figure 4: *Consonant-to-preceding vowel duration ratio as a function of consonant length and consonant type.*

Similarly, a significant interaction between consonant length and type was found in the ratio of consonant to following vowel duration (C/V2): $F(2, 426) = 5.273$; $p = 0.022$. Singletons differed from geminates in this parameter both for nasals and plosives ($p < 0.001$ in both cases). The average

values were higher in case of geminates compared to singletons (Figure 5).

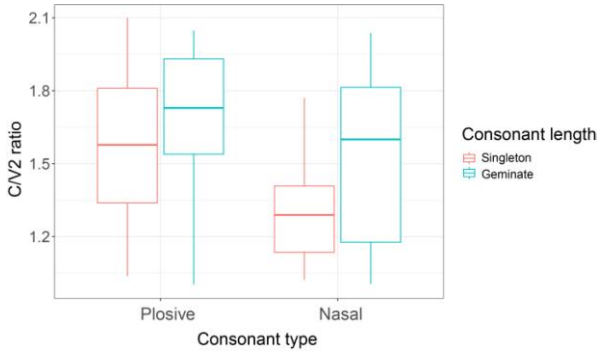


Figure 5: *Consonant-to-following vowel duration ratio as a function of consonant length and consonant type.*

In the next step, decision trees were applied to evaluate the contribution of each feature in reducing uncertainty associated with the target variables, thus aiding in the discrimination of the two quantity categories. The results show that for plosives, the two categories were distinguished primarily by consonant duration, while for nasals, the duration of the surrounding vowels also played an important role in addition to the consonant duration (Figure 6). Of the two relative durations, C/V2 seemed to be one of the more important features for both consonant types. C/V1 is less distinctive between the two quantity categories and may play a role more in plosives.

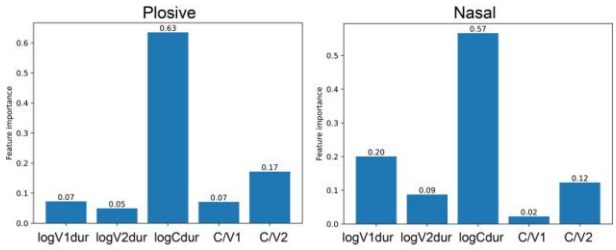


Figure 6: *Feature importance in distinguishing singletons and geminates in plosives and nasals.*

4. Discussion and conclusion

This study examined how the phonological quantity contrast of different consonant types is reflected in phonetic data. Our hypothesis was confirmed by the data, showing that speakers mark the contrast producing different durational patterns depending on the consonant type (nasal or plosive).

Our findings suggest that the expression of the quantity contrast in nasals requires more robust time adjustments than in plosives. In general, nasal consonant exhibited shorter durations in comparison to plosives. The relatively short durations may make the difference between short and long nasals less noticeable. Consequently, it is conceivable that supplementary features, such as the duration of surrounding vowels, play a role in marking the length contrast.

Results of the present study reflect 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. More specifically, there were differences in adjacent vowel durations depending on whether the following consonant was a nasal or a plosive. In Hungarian, the vowel preceding the target consonant (V1) seemed to be produced significantly longer before nasal geminates than before nasal singletons. However, this distinction was not evident with plosives. In future research, perceptual ratings on the data of the present corpus can help us establish whether the results found for Japanese can be applied to Hungarian.

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. To enhance our understanding of this phenomenon, in forthcoming investigations, we intend to conduct spectral analyses on the adjacent vowels.

5. 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]. Retrieved from http://www.praat.org

Deme, A., Bartók, M., Gráczi, T. E., Csapó, T. G., & Markó, A. (2019). Articulatory organization of geminates in Hungarian. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia (pp. 1739-1743).

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.

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82(13), 1–26.

Lahiri, A., Hankamer, J. 1988. The timing of geminate consonants. *Journal of Phonetics*, 16(3), 327–338.

Lenth, M. R. 2018. Package 'lsmeans'. *The American Statistician,* 34(4), 216–221.

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.

Neuberger, T. (2023). Geminate types and their acoustic effects on adjacent vowels in Hungarian. In Radek Skarnitzl, R. & and Jan Volín, J. (Eds.): *Proceedings of the 20th International Congress of Phonetic Sciences.* Guarant International.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Duchesnay E (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from https://www.r-project.org/

Ridouane, R. 2010. Geminates at the junction of phonetics and phonology. *Papers in laboratory phonology*, 10, 61–90.

Wickham, H. 2016. ggplot2 – *Elegant Graphics for Data Analysis* (2nd Edition). Springer-Verlag, New York.

Use of Natural Anchors for Dysarthria Assessment: An Exploratory Study on Improving Rater Reliability

Thushani Umesha Munasinghe[1], Deepthi Crasta[1], Kaila Stipancic[2], Mili Kuruvilla-Dugdale[1]
[1]*University of Iowa, Iowa, USA*
[2]*University at Buffalo, USA*
thushani-munasinghe@uiowa.edu, deepthi-crasta@uiowa.edu, klstip@buffalo.edu, mili-kuruvilla-dugdale@uiowa.edu,

Abstract

The aim of this project was to determine whether the use of anchors improves interrater and intrarater reliability when nonexpert listeners rated five features salient to hypokinetic dysarthria: overall severity, reduced loudness, articulatory imprecision, short rushes of speech, and monotony. Fourteen nonexperts rated 82 sentences recorded from individuals with Parkinson's disease and healthy controls using five separate equal appearing interval (EAI) scales to indicate their perception of the five features mentioned above. The listeners rated the samples twice, once without and once with external anchors. Interrater reliability and intrarater reliability were calculated using intraclass correlation coefficients (ICCs). Findings revealed an overall increase in both interrater and intrarater reliability for most features in the anchor condition, except for monotony, where a decrease in single-measures ICC was noted for the anchor compared to non-anchor condition. These preliminary findings highlight how external anchors can benefit interrater and interrater reliability when rating perceptual dimensions of dysarthria.

Keywords: scaling, dysarthria, reliability, anchors

1. Introduction

When listeners rate different speech samples, they develop internal standards for what counts as different intervals on a rating scale, and they rely on these standards to guide their judgements (Kreiman et al., 1993). Internal standards can be influenced by factors such as experience and training of the listener, or context, as well as other listener characteristics such as memory and attention (Gerratt et al., 1993; Kreiman et al., 1992). Internal standards are believed to be developed gradually over years, for example, when expert clinicians are exposed to disordered speech samples over time. Internal standards are known to drift and may be unstable while getting established, resulting in variable ratings. In addition to these listener-related factors, external factors such as acoustic context (e.g., a listener may perceive a speech sample with moderate severity as more severe if it is presented after a series of voices with mild dysarthria severity) and task (e.g., reading task vs. spontaneous conversation) can also affect internal standards.

The reliance on unstable internal standards for perceptual judgments could be one reason for the high variability in rater reliabilities reported in the literature. In the context of dysarthria, reduced rater reliability poses a significant challenge when employing auditory-perceptual assessments to diagnose and measure specific subsystem features (Stipancic et al., 2023). To improve rater reliability, voice researchers have used stable external standards with the intent of replacing the idiosyncratic internal standards (Awan & Lawson, 2009; Chan & Yiu, 2002). In these studies, the external anchor served as a reference against which raters could compare the experimental stimuli. Both natural (i.e., speech samples from speakers) and synthetic (i.e., computer-generated speech) anchors have been studied. Natural anchors seem to be the best references to use if they resemble the target stimuli to be rated by listeners.

Several studies have been conducted to examine the improvements to interrater and intrarater reliability when employing anchors in voice assessment, as well as the type of anchor (natural versus synthetic), and the mode of presentation (auditory, visual/textual, or a combination of both) that affects reliability most (Awan & Lawson, 2009; Santos et al., 2021). A previous study also explored how listener experience influenced the use of anchors (Eadie & Kapsner-Smith, 2011). The findings suggest that external auditory anchors enhance both interrater and intrarater reliability compared to other modalities. Experience level did not influence reliability, and both experienced and novice listeners demonstrated greater reliability when using anchors compared to the condition without anchors. Novice listeners often play a role in speech assessment, to determine the real-world impact of the communication disorder, and enables the recruitment of a large participant pool for research studies. However, no previous studies have been conducted to examine if and how rater reliability changes with external anchor use when nonexperts judge dysarthric speech features.

In the context of dysarthria assessment, interval scales are often used because they are less time consuming and easy to use in a clinical setting (Kreiman et al., 1993). The equal appearing interval scale (EAI) is one such scale, which has predefined intervals that are equidistant from each other. In an anchored condition, the experimenter provides a reference stimulus for each interval and raters can use the references to guide their ratings of the experimental stimuli. Although previous voice studies have explored the reliability of employing anchors with the EAI scale (Gerratt et al., 1993), similar investigations have yet to be conducted for dysarthria.

The aim of the present study was to compare the reliability of ratings completed with an EAI scale by nonexpert listeners, without and with the presence of anchors. Both interrater and intrarater reliability were examined for five salient hypokinetic dysarthria features, including overall speech impairment severity, reduced loudness, articulatory imprecision, short rushes of speech, and monotony.

2. Methods

This study was approved by the Institutional Review Board of the University of Iowa. All participants gave written informed consent before completing study procedures.

2.1 Participants

Two groups of participants were included: 1) speakers; and 2) listeners.

2.1.1 Speakers

The speakers consisted of 43 individuals with Parkinson's disease (PD; 18 females, 25 males) and 25 neurologically healthy speakers (11 females, 14 males). The inclusion criteria were: (i) no history of speech, language, or hearing disorders; (ii) no co-occurring neurological diagnosis for the participants with PD, and absence of any neurological diagnosis for the controls; (iii) no history of head and neck surgery; (iv) not wearing a hearing aid or having a prescription for hearing aids; and (v) be a monolingual, native speaker of American English.

2.1.2 Listeners

Fourteen neurologically healthy participants (M_{age} = 26.5 years, SD = 3.55) were recruited as listeners. The inclusion criteria were (i) be between the ages 19-90 years; (ii) pass a bilateral hearing screening at 25 dB HL at 500 Hz, 1 kHz, 2 kHz, and 4 kHz; (iii) no history of speech, language, or hearing disorders; (iv) use English as the primary language of communication; (v) have minimal exposure to communication disorders.

2.2 Experimental Tasks

2.2.1 Speech task

The speakers were recorded reading 11 unique sentences from The Speech Intelligibility Test (SIT; Yorkston et al., 2007); one sentence that presented with the greatest number of dysarthria features of interest was selected from each speaker. To determine which features were present, each sentence was rated on the Dysarthria Rating Scale (Darley et al., 1969 a, 1969 b), independently by two trained research assistants. Consensus was sought if they disagreed about features, and the consensus ratings were used to select the final stimulus set.

2.2.2 Auditory-perceptual scaling task

A total of 68 samples (i.e., 43 PD and 25 control) were used to determine interrater reliability; 20% of the samples (n = 14) were randomly selected and repeated for intrarater reliability.

The perceptual experiment was conducted in a quiet laboratory setting. Each listener attended two sessions which were one week apart and lasted approximately one hour each week. Listeners used calibrated headphones to listen to the speech samples. Ratings were completed using a custom MATLAB GUI that displayed five separate EAI scales at once, one for each feature (i.e., overall speech impairment severity, articulatory imprecision, reduced loudness, short rushes of speech, and monotony).

Definitions of each feature were provided by the experimenter to the listeners. They were instructed to rate overall severity based on their general impression of severity rather than understandability of the sentences. Reduced loudness was assessed based on the softness or quietness of the voice in the sample, while articulatory imprecision was evaluated based on how crisply and clearly the speech sounds were produced. Short rushes of speech was rated by identifying instances of rapid speech characterized as rushed segments preceded and followed by pauses. For monotony, the listeners were instructed to consider the flatness of the speech sample in terms of pitch, loudness, or duration.

Listeners used a 5-point EAI scale either without anchors or with anchors. The 5-point EAI scale had the following intervals: 1=typical, 2=mild, 3=moderate, 4=severe, and 5=profound. For the session without anchors, listeners were asked to rate the first three features after listening to a sample once and then rate the next two features after listening to the sample again. For the anchor condition, reference samples were provided for each scale interval for each feature. The listeners played the anchors of the first three features before listening to the sample and rating the three features. They followed the same steps for the next two features. The anchors reappeared after every eight samples. The listeners were encouraged to use the entire scale for the ratings during both the sessions.

The anchor for each scale interval was selected by two experts. First, an experienced speech-language pathologist rated dysarthria speech samples from the *Audio Seminar Series* (Darley et al., 1975) and chose samples for mild, moderate, severe, and profound levels for each feature. Then the last author rated the chosen samples independently for features and severity levels. Discrepancies between the experts were resolved through consensus before the anchors for each interval of each feature were selected.

2.3 Data Analysis

2.3.1 Statistical analysis

SPSS statistical software version 28 (SPSS Inc, Chicago, IL) was used for statistical analyses. Both interrater reliability and intrarater reliability were estimated using intraclass correlation coefficients (ICCs). For interrater reliability, single- and average-measures consistency from 2-way random-effects models (Koo & Li, 2016) with 14 raters across 68 samples was used to obtain the ICCs and their respective 95% confidence intervals. For intrarater reliability, single- and average-measures consistency from 2-way mixed-effects models (Koo & Li, 2016) for the 14 raters were calculated along with their relevant 95% confidence intervals.

Inter- and intra-rater reliability ICCs were calculated for each of the five speech features for both anchor conditions (i.e., with and without anchors). ICC values were descriptively compared between the anchor conditions for each feature. A difference in ICC values between anchor conditions was considered meaningful if there was a switch to a higher or lower reliability category with the use of anchors.

3. Results

3.1 Interrater reliability

Compared to the non-anchor condition, there was an overall increase in both single- and average-measures ICC for all features, including overall speech impairment severity, articulatory imprecision, reduced loudness, short rushes of speech, and monotony for the anchor condition (Table 1). The average-measures ICCs indicated good or excellent reliability regardless of the anchor condition. However, for short rushes of speech, a change in the reliability category was observed, where reliability increased from moderate to excellent with anchors. Single-measure ICC values for all features ranged from poor to moderate levels regardless of the anchor condition. However,

reliability for overall severity and reduced loudness improved from poor to moderate when anchors were used.

Table 1: *Interrater reliability of all features rated with an equal appearing interval (EAI) scale with and without anchors.*

Speech Feature	Interrater Reliability Intraclass Correlation Coefficient (CIs)		
	Measure	No Anchors	Anchors
Overall severity	Single	**0.492** (.403-.593)	**0.587** (.501-.680)
	Average	0.931 (.904-.953)	0.952 (.934-.967)
Articulatory imprecision	Single	0.513 (.425-.613)	0.597 (.512-.689)
	Average	0.937 (.912-.957)	0.954 (.936-.969)
Reduced loudness	Single	**0.423** (.337-.526)	**0.520** (.432-.619)
	Average	0.911 (.877-.940)	0.938 (.914-.958)
Short rushes of speech	Single	0.295 (.219-.393)	0.410 (.324-.513)
	Average	**0.854** (.797-.901)	**0.907** (.870-.936)
Monotony	Single	0.433 (.346-.536)	0.495 (.407-.596)
	Average	0.914 (.881-.942)	0.932 (.906-.954)

Note. CI=Confidence interval; 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. The bold values indicate the category shift in the ICC values.

3.2 Intrarater reliability

The single- and average-measures ICC values for intrarater reliability ranged from moderate to excellent in both anchor conditions (Table 2). Except for overall severity and monotony, reliability measures increased for both single and average measures when anchors were used. The single measure ICC for articulatory imprecision moved from moderate to good reliability, but there was no shift in reliability categories for any of the other features.

Although several tests are available to determine statistical differences between ICC measures (e.g., Fisher's Z test, Konishi-Gupta modified Z-test, the likelihood ratio test, and Alsawalmeh-Feldt F-test), for the present exploratory study, we compared the ICC measures in a more qualitative manner, as the study was underpowered (Donner et al., 2002).

4. Discussion and Conclusion

In this study, we investigated interrater reliability and intrarater reliability among raters who assessed speech samples from talkers with PD and healthy controls using an EAI scale. The preliminary findings presented here suggest that there are benefits to combining natural anchors with an interval scale for evaluating dysarthric speech.

A meaningful change in interrater reliability (i.e., switch to a higher reliability category) was observed for three out of the five speech features, namely overall severity (single-measures ICC), reduced loudness (single-measures ICC), and short rushes of speech (average-measures ICC). Despite this improvement, the moderate reliability observed for overall severity and reduced loudness when using anchors is insufficient for clinical purposes, which contrasts with the average measures for both anchor conditions, which are highly acceptable. Voice studies have reported similar magnitudes of improvements in reliability when using external anchors combined with training. However, these studies show increased interrater variability when anchors were used without training, suggesting limited use for anchors alone (Chan & Yiu, 2006). Most prior studies only include average-measures ICC, presumably because the individual rating is unreliable, and ICCs based on average measures are always higher than those based on single measures. Hayen and colleagues (2007) emphasize that average measures should not be used when determining ICCs unless there are specific situations where averaged ratings apply. In dysarthria assessment, the measurement from a single rater is typically the basis of the actual measurement, suggesting the importance of considering single measures ICCs.

Table 2: *Intrarater reliability of all features rated with an equal appearing interval (EAI) scale with and without anchors.*

Speech Feature	Intrarater Reliability Intraclass Correlation Coefficient (CIs)		
	Measure	No Anchors	Anchors
Overall severity	Single	0.824 (.773-.864)	0.824 (.773-.864)
	Average	0.903 (.872-.927)	0.903 (.872-.927)
Articulatory imprecision	Single	**0.734** (.663-.793)	**0.808** (.753-.852)
	Average	0.847 (.797-.884)	0.894 (.859-.920)
Reduced loudness	Single	0.761 (.695-.814)	0.813 (.759-.855)
	Average	0.864 (.820-.898)	0.897 (.863-.922)
Short rushes of speech	Single	0.650 (.561-.724)	0.719 (.643-.780)
	Average	0.788 (.719-.840)	0.836 (.783-.877)
Monotony	Single	0.734 (.662-.792)	0.668 (.583-.739)
	Average	0.847 (.797-.884)	0.801 (.736-.850)

The switch to a higher intrarater reliability category was observed only for single-measures ICC of articulatory imprecision. However, for most features, the magnitudes of both single and average measures increased with the use of anchors, except for monotony and overall severity. Similar improvements in intrarater reliability in the presence of auditory anchors have been observed in voice studies (Chan & Yiu, 2002; Eadie & Kapsner-Smith, 2011). Regarding monotony, previous dysarthria studies indicated that this feature behaves differently from other hypokinetic dysarthria features. Stipancic et al. (2023) showed that ratings of monotony had the poorest criterion validity and reliability compared to ratings of overall speech impairment, articulatory imprecision, and slow rate.

Another study by Stipancic (in press) identified monotony as a metathetic feature compared to the four other features in this study, which were identified as prothetic continua. Therefore, future work is necessary to identify the perceptual properties of monotony to delineate why it behaves differently, and to incorporate the findings for future research (e.g., have listeners rate the subordinate dimensions of monotone speech rather than overall monotony). When comparing single-measures ICC for interrater reliability and intrarater reliability, it was evident that the single-measures ICC for intrarater reliability were higher than for interrater reliability across anchor conditions and features. This indicates the raters are more consistent within themselves. In contrast, the average-measures ICC of intrarater reliability were lower than the interrater ICC values for both conditions. This contrasts with the findings of previous voice studies, which showed that intrarater reliability values are generally better than interrater reliability measures with the use of anchors (Awan & Lawson, 2009). The observed differences may stem from methodological variations across studies, including differences in subject populations, task complexities, and stimuli. Follow-up studies are warranted to investigate deeper into potential reasons underlying these disparities.

There may be potential reasons for the variability in interrater and intrarater reliability across different speech features. For the present study, we used the EAI scale to rate hypokinetic dysarthria features since it is one of the most used scales in research and clinical settings. However, an EAI scale might not be suited for assessing speech features that are prothetic because they are best scaled by direct magnitude estimation (DME), while metathetic features can be scaled using EAI or DME. Results of a recent study indicated that except for monotony, all the other features in the current study were prothetic, suggesting that a DME scale, rather than an EAI scale is the best fit to assess overall speech impairment severity, articulatory imprecision, reduced loudness, and short rushes of speech (Stipancic, in press). When selecting a scale for dysarthria assessment, it is essential to consider both reliability and validity. Even though the EAI scale shows increased reliability in rating dysarthria features with the use of anchors, it is also essential to consider if the EAI scale is the best fit for each feature. Critical next steps will be to examine both EAI and DME scales without and with anchors to see if reliability changes similarly across the different scales.

It is also important to consider when anchors should be used. A study by Stipancic et al. (2023) investigated the effect of auditory training on perceptual ratings of dysarthric speech, in which external anchors were only used during training and not during the post-training ratings. Results showed that there was little improvement in rater reliability as a result of training. One of the reasons might be that in this previous study, external anchors were only used during training, and the internal standards of the nonexpert listeners may have been unstable and insufficient on their own to improve reliability. Therefore, it is recommended to use anchors during the actual ratings to avoid overreliance on shifting or developing internal standards, particularly with nonexpert listeners.

It is important to systematically investigate the use of anchors for rating salient speech features of other dysarthria types. In the current study, an overall improvement in interrater and intrarater reliability was observed with the addition of external anchors to an EAI scale. The feasibility of incorporating anchors for other types of scales, such as DME and visual analog scales will be investigated in future work.

5. Acknowledgements

This research was supported by the National Institutes of Health (R15DC016383 and R21DC019952; PI: Kuruvilla-Dugdale). We are grateful to the research assistants and subjects who participated in the study. Special thanks to Dahlia Cukierkorn, Lexi Jacobsmeyer, Morgan Linneweh, Ella Meier, and Anna Mae Williams for helping with data collection and analysis. Chaewon Park and Minguang Song helped modify the original MATLAB GUI for the anchor condition.

6. 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: JSLHR*, *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.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.

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.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research*, *35*(3), 512–520.

Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*, *17*(1), 45–56.

Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, *40*(3), 699–704.

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.

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*, 1–23.

Stipancic, K. L., Whelan, B., Laur, L., Zhao, Y., Rohl, A., & Kuruvilla-Dugdale, M. (in press). Tipping the Scales: Indiscriminate Use of Interval Scales to Rate Diverse Dysarthric Features. *Journal of Speech, Language, and Hearing Research.*

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

Abstract

Keywords: Speech, acoustics, perception, naturalness, articulatory synthesis

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. 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. For this purpose, a perception experiment using a 4-points Likert scale was conducted to evaluate the naturalness of seven static phonemes, /a, e, i, ə, f, s, ʃ/, synthesized with a 1D and a 3D models. No significant influence of the acoustic model was found, however, significant differences between the phonemes were perceived.

1. Introduction

Articulatory synthesis relies on a description of the physical phenomena involved in speech production. It uses a geometrical description of the speech production apparatus and models the sound generation and propagation mechanisms.

A very common simplifying assumption is to consider that the acoustic propagation is unidimensional, i.e. it depends only on the cross-sectional area along the vocal tract (Sondhi and Schroeter 1987). However, this assumption is increasingly unrealistic toward HF. The divergence with realistic models first appears as shifts in resonance frequencies due to the curvature of the acoustic field at changes in cross-sectional area. At HF, above about 4-5 kHz, the higher order modes generate additional resonances unpredicted by 1D models (Blandin, Arnela, Laboissière, et al. 2015). These phenomena can be properly described by 3D models, such as finite elements (Arnela et al. 2019), finite differences (Takemoto, Mokhtari, and Kitamura 2010), the multimodal method (Blandin, Arnela, Félix, et al. 2022) or waveguide mesh models (Gully, Daffern, and Murphy 2017).

So far, articulatory synthesis based on 3D acoustic models has been achieved for isolated phonemes (Gully, Daffern, and Murphy 2017; Arnela et al. 2019; Dabbaghchian et al. 2021). One can expect that using more realistic acoustic models for articulatory synthesis would result in a greater resemblance to actual human speech, and that it would be perceived as more natural. However, the hearing sensitivity toward HF reduces both in terms of sound pressure level (SPL) and frequency discrimination. Thus, this increase of realism, which happens mostly at HF, may not substantially impact the perceived naturalness. This implies the necessity to evaluate the perceptual impact of such models.

Prior to our study, to our knowledge, only one study addressed this question using a perceptual test. 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 3D simulation method used, waveguide mesh, is non standard and not very well proven, so the increase of realism can be questioned. 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 mechanisms is necessary.

Our objective was to determine whether an articulatory synthesis based on a 3D acoustic model with a frequency domain approach results in phonemes perceived as more natural.

To that end, four vowels (/a, i, u/ and /ə/) and three consonants (/f, s, ʃ/) were synthesized for a male and a female speaker. For this purpose, we applied a source-filter approach in which the filter (vocal tract transfer function) was computed with both 1D and 3D acoustic models.

2. Methods

2.1. Stimuli generation

The stimuli were generated with the articulatory synthesizer VocalTractLab3D[1] (Blandin, Arnela, Félix, et al. 2022), which can synthesize speech sounds with a 1D or a 3D acoustic model. The vocal tract geometries used are predefined in VocalTractLab3D. They have been generated by fitting the parameters of the geometric vocal tract model implemented in VocalTractLab3D to magnetic resonance images (MRI) obtained for multiple phonemes produced by a male (Birkholz 2013) and a female (Drechsel et al. 2019) speaker.

The 3D simulation method implemented in VocalTractLab3D is a multimodal method which relies on a decomposition of the acoustic field $p(x, y, z)$ over the local transverse modes $\varphi_n(y, z)$:

$$p(x, y, z) = \sum_{n=0}^{\infty} p_n(x)\varphi_n(y, z), \quad (1)$$

where $p_n(x)$ descibes the amplitude of the transverse mode $\varphi_n(y, z)$ along the vocal tract.

A complete description of the method can be found in Blandin, Arnela, Félix, et al. 2022. Its main advantages are to be computationally efficient and to provide a better understanding of the physical phenomena involved. In the context of our study, another advantage is the possibility to tune the dimension of the model through the number of transverse modes used: using only one transverse modes makes a 1D simulation and using a correctly tuned number makes a 3D simulation. This tuning was done through convergence tests and comparison with finite elements simulations (Blandin, Arnela, Félix, et al. 2022).

Several vocal tract transfer functions were computed:

- for the vowels (/a, i, u, ə/), from the volume velocity at the glottis and from the acoustic pressure at a point about 2 cm downstream of the glottis to the acoustic pressure at a point located 1 m in front of the lips,
- for the fricatives (/f, s, ʃ/), from the acoustic pressure at a point in the sound generation area (teeth or hard palate) to the acoustic pressure at a point located 1 m in front of the lips. This point source was placed between the lips for /f/, at the downstream edge of the lower lips for /s/, and between the teeth for /ʃ/. Its location was fine tuned to reproduce properly the intended phonemes.

The vocal fold sound source signal was generated using the Liljencrants- Fant (LF) glottal pulse model (Fant, Liljencrants, Lin, et al. 1985) implemented in VocalTractLab3D. The fundamental frequency was set to a target of 120 Hz and 210 Hz for the male and female voices, respectively. To increase the naturalness of the stimuli, small variations of fundamental frequency were generated with a "flutter" as proposed in Eq. (1) in Klatt and Klatt (D. Klatt and L. Klatt 1990). An open quotient of 0.5, a shape quotient of 3.0 and spectral tilt of 0.02 were used in order to generate a modal voice quality which corresponds to normal speech.

The noise sources present immediately downstream of the vocal folds for the vowels and in the vicinity of obstacles for the fricatives were generated by filtering Gaussian white noise with a first-order low-pass filter. Cut-off frequencies of 10 kHz for the vowels, 5 kHz for /f/, and 8 kHz for /s, ʃ/ were used. These values roughly create source spectra according to Shadle 1991. The gain of the sources was adjusted in such a way that the intensity of the produced noise at the different places of articulation closely matches real fricative intensities (Birkholz 2014).

To generate the stimuli, the source signals were convolved with the impulse responses of the transfer functions. In the case of the vowels, the amplitude p_s of the noise source was set proportional to the cube of the low frequency part of the vocal fold output particle velocity $\bar{v}$, $p_s \propto |\bar{v}^3|$ as proposed by Stevens (Stevens 2000). Applying the principle of superposition of linear acoustics, the signals from the noise source attenuated by 30 dB and the vocal fold were then added to form the radiated sound. In total, 28 stimuli were generated: 2 acoustic models (1D or 3D)×7 phonemes×2 genders.

2.2. Perception experiment

Naturalness was evaluated by 31 participants aged between 21 and 28 years old (4 males and 27 females), all native French speakers without past or present hearing problems. They all had hearing thresholds $\leq$ 20 dB hearing level (HL) bilaterally at octave frequencies between 500 and 8000 Hz (audiometric screening with pure-tone audiometry using a MADSEN Itera II audiometer with TDH-39 earphones). The experiment took place in a listening booth where the stimuli were played through a loudspeaker placed one meter in front of the participants. The choice of a loudspeaker instead of headphones was motivated by the better control over the listening conditions that it offers and the fact that it is closer to a real life listening condition. In addition, it eliminates the problem of achieving the same HF response for all participants, which is challenging with headphones. The gain of the amplifier of the loudspeaker was adjusted so that the level of the stimuli at the location of the head of the participants was 70 dB SPL. Participants listened to each stimulus as many times as they wanted and were asked to rate it on a 4-points 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.

2.3. Statistical analysis

Participants' responses were analyzed with an ordinal cumulative logistic regression model using the "ordinal" R packages (Christensen 2015). A random effect of the participant was used and the fixed effects were the acoustic model (two conditions: the 1D and 3D models), the type of phoneme (/a, i, u, ə, f, s, ʃ/), the gender of the speaker (female and male) and the moment of the test (two moments: test and retest). The model included each main factor, the interactions between the model and the phoneme, and the interaction between the model and the gender. 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). The Holm method of alpha adjustment was used to correct for multiple testing. Inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC) (Shrout and Fleiss 1979).

3. Results

Figure 1 shows the average rating for each phoneme synthesized with both acoustic models. The level of inter-rater relia-

[1] VocalTractLab3D is freely available at: https://vocaltractlab.de/index.php?page=vocaltractlab-download

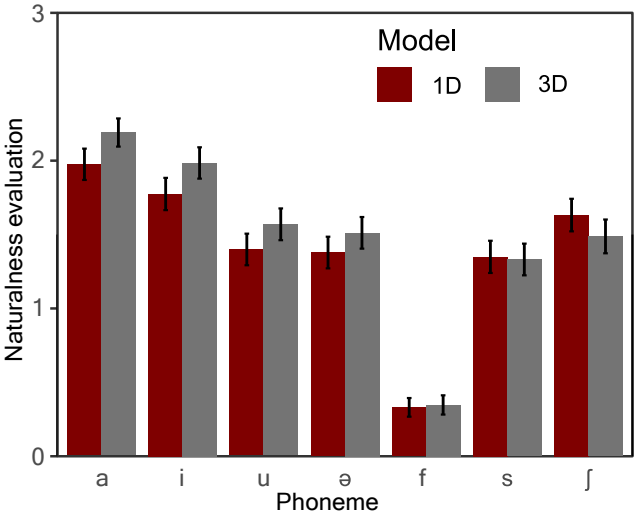


Figure 1: Average ratings for the phonemes synthesized with the 1D and 3D acoustic models in the naturalness rating task using a Likert scale from 0 (not at all natural) to 3 (completely natural).

bility can be regarded as good to excellent with ICC = 0.9 (with 95% confident interval = 0.86 - 0.94 and p < .0001). 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). The interaction between the model and the gender was non-significant (χ^2(1) = 0.021, p = 0.885), as well as the interaction between the model and the phoneme (χ^2(6) = 6.82, p = 0.337). However, a significant effect of the phoneme was found (χ^2 (6) = 464, p < 0.001).

As depicted in Fig. 1, 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".

4. Discussion and conclusion

In contrast to Gully (2017), our results do not show a significant influence of the 3D acoustic model on the perceived naturalness. This discrepancy between the two studies could be explained by differences in the simulation method, the phonetic material (isolated phonemes including consonants vs. diphthongs), the listening conditions (loudspeaker vs. headphones), or the experimental design (Likert scale vs. MUSHRA (Series 2014)). Additionally, the use of electrolaryngograph signals from human subjects for the sound source in the study of Gully might generate globally more natural sounding stimuli than the LF model.

In Fig. 1, the average naturalness of the vowels is slightly better for the 3D model compared to the 1D model. On the other hand, the p-value of the effect of the model (p = 0.085) is close to 0.05, which is the usual limit to consider an effect as significant. This suggests that a weak but significant effect might could be revealed using more participants, and/or different experimental design choices, such as a linear scale instead of a Likert scale. This tends to be confirmed in a subsequent study by Blandin, Stone, et al. 2023, showing significant differences using pair comparisons between 1D and 3D models, and a linear scale to rate the naturalness. However, only 5 vowels (/a, e, i, o, u/) were used and the frequencies up to 4 kHz were similar for each model. The perceived differences between 1D and 3D mainly concern the vowels /o/ and /u/.

As shown in Fig. 1, 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).

Regardless of the acoustic model, there are significant differences of naturalness between the phonemes. This confirms that the perceptual experiment was successful in detecting variations of naturalness, but that the effect of the model, if existent, is probably too small to be observed this way. On the other hand, this also means that other phoneme-specific factors have more impact than the acoustic model.

Given the multiplicity of the phenomena involved, it is difficult to identify accurately which phenomenon is affecting naturalness the most for a specific phoneme. However, one can formulate hypotheses. For example, the sound generation is expected to take place in the vicinity of the lips for /f/. Therefore, the simplification of the lip shape as a flat opening may degrade the naturalness more for this specific phoneme. This may explain the particularly low rating for /f/. More generally, other causes may negatively affect the naturalness of the synthetic fricatives. The simplification of the aeroacoustic sound sources as a single point source may be a too rough approximation, 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 further degraded by the radiation simplifications.

Regarding the vowels, the source filter coupling (Titze 2008) was not taken into account in this study. The dependence of this phenomenon on the vocal tract shape may contribute to differences of the naturalness between the vowels (Birkholz et al. 2019): for vowels having a greater source filter coupling, not taking it into account may affect more their naturalness. This is in line with the results of Birkholz et al. 2019 who reported a stronger effect on close-mid to close vowels (/i, ə, u/) for which a lower naturalness was observed. In addition, the participants are not used to listening to the vowel /ə/ in isolation in natural speech. This may explain why it has the lowest naturalness among the vowels.

5. Acknowledgements

This study was supported by the German Federal Ministry of Education and Research (BMBF) within the project "Promise-AI", funding code 16SV8988.

6. References

Arnela, M, S Dabbaghchian, O Guasch, and O Engwall (2019). "MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs". In: *IEEE Trans. Audio Speech Lang. Process.* 27.12, pp. 2173–2182.

Birkholz, P (2013). "Modeling consonant-vowel coarticulation for articulatory speech synthesis". In: *PloS one* 8.4, e60603.

— (2014). "Enhanced area functions for noise source modeling in the vocal tract". In: *10th International Seminar on Speech Production, Köln*, pp. 1–4.

Birkholz, P, F Gabriel, S Kürbis, and M Echternach (2019). "How the peak glottal area affects linear predictive coding-based formant estimates of vowels". In: *J. Acoust. Soc. Am.* 146.1, pp. 223–232.

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, M Arnela, R Laboissière, X Pelorson, O Guasch, A Van Hirtum, and X Laval (2015). "Effects of higher order propagation modes in vocal tract like geometries". In: *J. Acoust. Soc. Am.* 137.2, pp. 832–843.

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 Trans. Audio Speech Lang. Process.*

Christensen, RHB (2015). *Ordinal—regression models for ordinal data, 2015. R package version 2015.6-28*.

Dabbaghchian, S, M Arnela, O Engwall, and O Guasch (2021). "Simulation of vowel-vowel utterances using a 3D biomechanical-acoustic model". In: *Int. J. Numer. Methods Biomed. Eng.* 37.1, e3407.

Drechsel, S, Y Gao, J Frahm, and P Birkholz (2019). "Modell einer Frauenstimme für die artikulatorische Sprachsynthese mit VocalTractLab". In: *Konferenz Elektronische Sprachsignalverarbeitung*. TUDpress, Dresden, pp. 239–246.

Fant, G, J Liljencrants, Q Lin, et al. (1985). "A four-parameter model of glottal flow". In: *STL-QPSR* 4.1985, pp. 1–13.

Gully, AJ (2017). "Diphthong Synthesis using the Three-Dimensional Dynamic Digital Waveguide Mesh". PhD thesis. University of York.

Gully, AJ, H Daffern, and DT Murphy (2017). "Diphthong synthesis using the dynamic 3D digital waveguide mesh". In: *IEEE/ACM Trans. Audio, Speech, Language Process.* 26.2, pp. 243–255.

Jiang, J and T Nguyen (2007). *Linear and generalized linear mixed models and their applications*. Vol. 1. Springer.

Klatt, DH and LC Klatt (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers". In: *J. Acoust. Soc. Am.* 87.2, pp. 820–857.

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*.

Shadle, CH (1991). "The effect of geometry on source mechanisms of fricative consonants". In: *Journal of phonetics* 19.3-4, pp. 409–424.

Shrout, PE and JL Fleiss (1979). "Intraclass correlations: uses in assessing rater reliability." In: *Psychological bulletin* 86.2, p. 420.

Sondhi, M and J Schroeter (1987). "A hybrid time-frequency domain articulatory speech synthesizer". In: *IEEE/ACM Trans. Audio, Speech, Language Process.* 35.7, pp. 955–967.

Stevens, KN (2000). *Acoustic phonetics*. Vol. 30. MIT press.

Takemoto, H, P Mokhtari, and T Kitamura (2010). "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method". In: *J. Acoust. Soc. Am.* 128.6, pp. 3724–3738.

Titze, IR (2008). "Nonlinear source–filter coupling in phonation: Theory". In: *J. Acoust. Soc. Am.* 123.5, pp. 2733–2749.

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]

Abstract

The somatosensory effect of electromagnetic articulography (EMA) sensors on speech remains relatively unexplored. Moreover, EMA sensors may be more disruptive to speech in individuals with somatosensory deficits (e.g., persons with Parkinson's Disease; PwPD). Thus, we investigated the effect of EMA sensors on the articulatory-acoustic vowel space (AAVS) in both typical speakers (n=23) and PwPD (n=23). The AAVS was calculated before EMA sensor placement, directly after, and after approximately one hour to assess habituation. The AAVS significantly decreased following sensor placement and did not change with habituation, regardless of speaker group. PwPD had a smaller AAVS compared to typical speakers, but were not differentially impacted by EMA sensors. EMA sensor placement led to average reductions of the AAVS of 13.5% for PwPD and 14.2% for typical speakers, which suggests that articulatory-acoustics from studies with and without the use of EMA sensors may not be fully comparable.

Keywords: Electromagnetic articulography, speech acoustics, Parkinson's Disease

1. Introduction

Electromagnetic articulography (EMA) provides fine-grained spatial and temporal information on articulatory movements during speech. While the primary outcome measures of speech studies using EMA are kinematic trajectories, it is not uncommon to also collect parallel acoustic data (Mefferd and Green 2010; Lee, Littlejohn, and Simmons 2017; Thompson and Kim 2019). However, when using EMA, the sensors that are attached to the tongue, jaw and lips may alter the speaker's articulatory-acoustic output as they might interfere with one's articulation. This raises the question to what extent the articulatory-acoustic output with EMA sensors on represents the typical output of a speaker. Given that the sensor coils also change the somatosensory feedback a speaker receives, it further raises the question as to whether the presence of EMA sensors impacts the articulatory-acoustic output of those with sensory deficits, such as persons with Parkinson's Disease (PwPD), to a greater extent than typical speakers (Conte et al. 2013). Parkinson's disease (PD) is a progressive neurodegenerative disease that affects various aspects of motor and sensory functioning, including the speech subsystems (Opara et al. 2017; Broadfoot et al. 2019; Chen and Watson 2017).

Previous studies assessing the impact of EMA sensors on articulatory-acoustics yielded mixed findings across various speaker populations, including typical speakers (Dromey, Hunter, and Nissen 2018; Bartholomew 2020), individuals with apraxia of speech (AOS; Katz, Bharadwaj, and Stettler 2006), and PwPD (Hirsch, Thompson, and Kim 2024). Katz, Bharadwaj, and Stettler (2006) showed that EMA sensors did not cause consistent group-level articulatory-acoustic effects on the production of vowels and fricatives in target words produced by individuals with and without AOS. In contrast, Dromey, Hunter, and Nissen (2018) showed that following sensor placement, the centre of gravity of sibilants embedded in target words was significantly reduced and did not increase over the course of habituation (20 minutes) in typical speakers. Moreover, Bartholomew (2020) observed a decrease in the first formant frequency (F_1) in target words four minutes after sensor placement compared to directly after EMA sensor placement for typical speakers, but comparisons to a pre-placement baseline were not conducted. Lastly, Hirsch, Thompson, and Kim (2024) reported a lower centre of gravity in sibilants directly after EMA sensor placement in speakers with and without PD compared to before sensor placement using a reading passage. In the same study and passage, the authors reported no significant differences in the quadrilateral vowel space area (q-VSA) for speakers with and without PD. However, Hirsch, Thompson, and Kim (2024) did not assess habituation to the sensors over a longer period of time between individuals with and without PD. Thus, the question remains as to what extent the presence of sensor coils across a longer time period may differentially affect PwPD, also in terms of habituation to the sensors themselves.

Therefore, the objective of this study was to determine the effect of EMA sensors on a sentence-level articulatory-acoustic measure of speech for both typical speakers and PwPD. We also assessed whether speakers habituated over time (approximately 60 minutes) and whether habituation varied by speaker group. If speakers adapt to the somatosensory changes introduced by the sensor coils, we would expect the AAVS after a long period of habituation to be significantly larger than the AAVS directly after sensor placement and comparable to the AAVS prior to the sensor placement. If speakers do not adapt to the EMA sensors, we would expect no significant differences in AAVS as a function of time since sensor placement.

2. Methods

2.1. Participants

This study used data from a previous study that received ethical clearance from the institutional Medical Ethics Review Board (NL66063.042.18; Jacobi 2022). We used the data from 46 individuals who gave written permission for their data to be used for follow-up studies. This included 23 typical speakers (18 male, 5 female; mean age = 68.4 years, standard deviation (SD) = 6.2) and 23 PwPD (18 male, 5 female; mean age = 69.1 years, SD = 7.0). Four other speakers participated, but were excluded as they either did not have recordings before sensor placement

(n=3) or were not diagnosed with idiopathic PD (n=1). Speakers did not report any hearing, speech, or neurological problems (other than PD) through self-report. All participants were native speakers of Dutch. PwPD participated while ON levodopa and had been diagnosed with idiopathic Parkinson's disease by a neurologist one to 19 years prior to their participation in the study.

2.2. Procedures

All speakers read the Dutch version of the North Wind and the Sun passage before and after EMA sensors (Northern Digital Inc. Wave system) were attached to the tongue, jaw, and lips (Jacobi 2022). Two sensors were placed on the tongue: one approximately one cm from the anatomical tongue tip, and one five mm anterior of the participant's /k/ constriction. Sensors were also placed on the jaw, and the vermilion border of the upper and lower lips. Acoustic data were assessed at three time points: time point 0 (T0), prior to sensor placement; time point 1 (T1), directly after sensor placement; and time point 2 (T2), at the end of the experiment, which lasted approximately one hour and consisted of multiple speaking tasks. The T2 recording was only made for 32 speakers, including 14 PwPD (10 male, 4 female), and 18 typical speakers (14 male, 4 female). Speakers were recorded in a quiet room of their own home with a microphone (Audio Technica AT875R) at a 22,050 Hz sample rate with a mouth-to-mic distance of approximately 20 cm.

2.3. Acoustic analysis

Any speech segments from the researcher giving instructions or any loud background noise (e.g., a clock) were removed from the speech recordings. All voiceless segments were subsequently removed from the speech recordings using a custom script in Praat 6.3.1 (Boersma and Weenink 2023). From these voiced segments, continuous first and second formant (F_1 and F_2) traces were extracted in Praat using a script based on Carignan (2022). As Escudero et al. (2009) showed, formant tracking accuracy is heavily influenced by both speaker and vowel characteristics. The Carignan (2022) script therefore aims to calculate the 'optimal' formant value by extracting the F_1 and F_2 with formant ceilings ranging from 3,500-6,000 Hz with 50 Hz intervals, resulting in 51 measurements (one for each ceiling) per analysis frame. The script uses the Burg algorithm, time steps of 5 ms, and a 25 ms time window. From these 51 possible formant values, those two standard deviations away from the mean formant value were removed. From the remaining formant values, the median value was taken as the optimal formant frequency of a particular 5 ms time step.

The resulting formant traces were filtered using a median absolute deviation filter which removed data points 2.5 times away from the median absolute deviation of the dataset. This removed 16,626 rows (4.1%), where every row corresponds to a 5 ms time step.[1] The AAVS was calculated on a mel-scale based on these filtered trajectories per speaker and time point, resulting in two or three AAVS values per speaker depending on whether the T2 recording was made. To calculate the AAVS, we used the methods established in earlier work (Whitfield and Goberman 2014; Abur, Perkell, and Stepp 2022). First, we computed the squared variance of both the F_1 and F_2 tracks. Next, we calculated the unshared variance by subtracting the R^2 of a linear model with F_1 predicting F_2 from 1. Finally, we take the square-root of the product of the squared variance and unshared variance (see Formula 1).

$$AAVS = \sqrt{(\sigma_{F_1})^2 \times (\sigma_{F_2})^2 \times (1 - R^2)} \quad (1)$$

[1] Additional manual filtering removed an extra 605 rows (0.2%). The results with and without manual filtering were nearly identical and we therefore use the AAVS with the median absolute deviation filter only.

2.4. Statistical analysis

Linear mixed-effects models were used to analyse the data in R 4.3.2 (R Core Team 2023; Bates et al. 2015; Kuznetsova, Brockhoff, and Christensen 2017). Our hypothesis model included the effect of group (PwPD vs. Typical) and time (T0, T1, T2) on the AAVS, and a by-speaker random intercept. All numerical variables were centered around the mean. We assessed whether adding an interaction between group and time improved the fit of the model by using the *anova()* function. A p-value below .05 would indicate that the interaction significantly improves the model.

Following our hypothesis test, we assessed the effects of speaker sex and age in an exploratory manner using model selection procedures, as these variables may impact vowel formants. We compared models using the *anova()* function and kept the more complex level if it significantly improved the fit of the model (i.e., $p < .05$).

To conclude our analysis, we employed model criticism by refitting our model on a trimmed dataset in which we removed data points whose residuals were at least two SDs away from their fitted value (Baayen 2008, Chapter 6). We used this trimmed data set if, and only if, outliers drove the presence or absence of any significant effects. Finally, we verified that the model met the assumptions of normality, homoscedasticity, multicollinearity and autocorrelation (Fox and Weisberg 2019).

3. Results

Our results are based on the dataset with trimmed residuals in which 5 data points (4.03%) were removed. Descriptive results per sex, group, and time are provided in Figure 1A. The AAVS at T0 was significantly larger compared to T1 (β = 3,325 mel^2, T = 8.0, CI = [2,433, 4,096], $p < .001$). On average, the AAVS was 13.5% smaller for PwPD at T1 compared to T0, and 14.2% smaller for typical speakers. There was no significant difference between the AAVS at T2 and T1 ($p = .30$). On average, the AAVS was 0.7% larger at T2 compared to T1 for PwPD, and 5.6% larger for typical speakers. A main effect of group indicated that PwPD had a significantly smaller AAVS compared to typical speakers overall (β = -5,850 mel^2, T = -4.9, CI = [-8,035, -3,577], $p < .001$). The addition of an interaction between time and group did not improve the fit of the model ($\chi^2(2) = 1.14$, $p = .57$, see Figure 1B).

Our subsequent exploratory analysis revealed a significant effect of sex which indicated that males had a lower AAVS compared to females (β = -10,101 mel^2, T = -7.0, CI = [-13,055, -7296], $p < .001$). Secondly, a significant effect of age indicated that AAVS decreased with speaker age (β = -191 mel^2, T = -2.1, CI = [-375, -22], $p = .04$). The inclusion of the exploratory variables did not alter the significance levels of the terms included in our hypothesis model.

4. Discussion and conclusion

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 per-

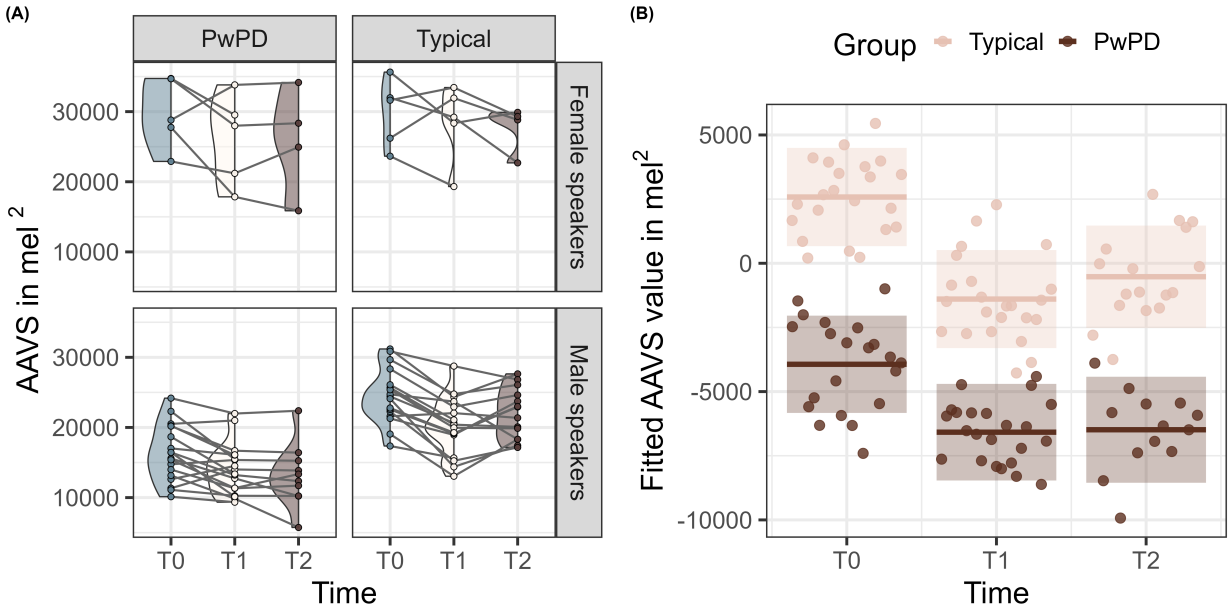


Figure 1: **(A)** Articulatory-acoustic vowel space (AAVS) per time point (T0, T1, T2) and group (Typical, PwPD), by sex (Male, Female). Different colours represent different time points (T0: blue, T1: white, T3: brown). Each point represents an individual speaker. **(B)** Model output showing the fitted mean-centered AAVS values of each group (Typical speakers; cream, Persons with Parkinson's disease: brown) per time point.

sons with Parkinson's Disease (PwPD). The results suggest that the AAVS is reduced after EMA sensor placement and does not significantly increase with habituation regardless of speaker group. This is in line with the results reported by Dromey, Hunter, and Nissen (2018), who previously reported no significant acoustic adaptation for /s/ and /ʃ/.

We did not find evidence that PwPD are affected by the EMA sensors to a different extent than typical speakers, which suggests that group differences in AAVS were not impacted due to the placement of EMA sensors. Our results are consistent with prior work that showed comparable EMA sensor effects on sibilants between speakers with and without dysarthria, and extend the findings from sibilants and individual vowel formant metrics to sentence-level vowel metrics computed over running speech (Katz, Bharadwaj, and Stettler 2006; Hirsch, Thompson, and Kim 2024). Our results underscore the reliability of using EMA in assessing speech motor functions in PwPD despite possible sensory integration changes that arise as a consequence of PD (Conte et al. 2013). PwPD did have an overall lower AAVS than typical speakers when accounting for sex and age differences, which is in line with previous work (Skodda, Visser, and Schlegel 2011; Whitfield and Goberman 2014; Tjaden, Lam, and Wilding 2013).

The results further 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, with average reductions of 13.5% for PwPD and 14.2% for typical speakers. This contrasts with Katz, Bharadwaj, and Stettler (2006), who reported no significant change in F_1 and F_2 measured with and without EMA sensors. However, the task also differed: we employed a reading passage whereas Katz, Bharadwaj, and Stettler (2006) used target words embedded in a carrier phrase, which might have elicited more clear speech. Our results further contrasts with those reported by Hirsch, Thompson, and Kim (2024) as they also did not report statistically significant reductions of the q-VSA following EMA sensor placement compared to pre sensor placement. One possible explanation for the difference is that the AAVS takes all vowels into account and provides an indication of general working space (i.e., the size of the space speakers tend to use the most), whereas previous studies looked at individual vowel formants or vowel formant metrics that provide more absolute indications of the vowel space (i.e., the maximum size of the vowel space).

Lastly, our results showed an effect of age such that the AAVS decreased with speaker age in our sample (age range: 52-81 years), regardless of speaker group. This finding might be explained by age-related atrophy of the orofacial and tongue musculature, which might result in smaller articulatory movements (Neel and Palmer 2012). However, it is important to note that the effects of aging on the size of the vowel space have been inconsistent, and that we did not test any young or middle aged adults (see e.g., Kent and Vorperian 2018; Hermes, Audibert, and Bourbon 2023).

A limitation of our study was that we could only assess habituation at the end of the experiment for a subset of participants (32/46 speakers). Considering that speakers were tested at home, the different locations may have resulted in different levels of background noise. To account for this, we checked the acoustic recordings and ensured an appropriate signal to noise ratio was present for all recordings (> 30 dB; Deliyski, Shaw,

and Evans 2005), and levels ranged from 33.6-59.4 dB (mean: 44.5 dB).

In conclusion, we show that passage-level vowel formant metrics are reduced as a result of EMA sensor placement, with an average reduction of 13.5% for PwPD and 14.2% for typical speakers. The AAVS did not increase after a long period of habituation regardless of speaker group. As a result, articulatory-acoustic vowel metrics from studies with and without parallel EMA data acquisition might not be comparable. Moreover, our results show that individuals with and without PD are impacted by the presence of EMA sensors in a similar manner, underscoring its reliability in assessing the speech motor functions in PwPD.

5. Acknowledgements

We are most grateful to all speakers included in the original study who kindly agreed to have their data be used for scientific purposes. This work was supported by a University of Groningen Center for Language and Cognition PhD grant awarded to the first author and by the Research School of Behavioral and Cognitive Neurosciences of the University of Groningen. This work was further supported by the International Macquarie University Research Excellence Scholarship (iMQRES) grant awarded to the third author.

6. References

Abur, Defne, Joseph S. Perkell, and Cara E. Stepp (2022). "Impact of Vocal Effort on Respiratory and Articulatory Kinematics". In: *Journal of Speech, Language, and Hearing Research* 65.1, pp. 5–21.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

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.

Boersma, Paul and David Weenink (2023). *Praat: doing Phonetics by computer [computer programme]*. Version Number: 6.3.1. URL: https://praat.org.

Broadfoot, C. K., D. Abur, J. D. Hoffmeister, C. E. Stepp, and M. R. Ciucci (2019). "Research-Based Updates in Swallowing and Communication Dysfunction in Parkinson Disease: Implications for Evaluation and Management". In: *Perspectives of the ASHA Special Interest Groups* 4.5, pp. 825–841.

Carignan, Christopher (2022). *Formant Optimization*. URL: https://github.com/ChristopherCarignan/formant-optimization.

Chen, Yu-Wen and Peter J. Watson (2017). "Speech production and sensory impairment in mild Parkinson's disease". In: *The Journal of the Acoustical Society of America* 141.5, pp. 3030–3041.

Conte, Antonella, Nashaba Khan, Giovanni Defazio, John C. Rothwell, and Alfredo Berardelli (2013). "Pathophysiology of somatosensory abnormalities in Parkinson disease". In: *Nature Reviews Neurology* 9.12, pp. 687–697.

Deliyski, Dimitar D., Heather S. Shaw, and Maegan K. Evans (2005). "Adverse Effects of Environmental Noise on Acoustic Voice Quality Measurements". In: *Journal of Voice* 19.1, pp. 15–28.

Dromey, Christopher, Elise Hunter, and Shawn L. Nissen (2018). "Speech Adaptation to Kinematic Recording Sensors: Perceptual and Acoustic Findings". In: *Journal of Speech, Language, and Hearing Research* 61.3, pp. 593–603.

Escudero, Paola, Paul Boersma, Andréia Schurt Rauber, and Ricardo A. H. Bion (2009). "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese". In: *The Journal of the Acoustical Society of America* 126.3, pp. 1379–1393.

Fox, John and Sanford Weisberg (2019). *An R Companion to Applied Regression*. Third edition. Thousand Oaks CA: Sage.

Hermes, Anne, Nicolas Audibert, and Angélina Bourbon (2023). "Age-related vowel variation in French". In: *Proceedings of the 20th International Congress of Phonetic Sciences*. Prague, Czech Republic, pp. 2045–2049.

Hirsch, Micah E., Austin Thompson, and Yunjung Kim (2024). "The Effects of EMA Sensors on Speech in Individuals with and without Dysarthria". In: Poster presented at the 22nd biennial Motor Speech Conference, San Diego, CA, United States of America.

Jacobi, Jidde (2022). "Coordination and timing of speech gestures in Parkinson's disease". PhD thesis. University of Groningen.

Katz, William F., Sneha V. Bharadwaj, and Monica P. Stettler (2006). "Influences of Electromagnetic Articulography Sensors on Speech Produced by Healthy Adults and Individuals With Aphasia and Apraxia". In: *Journal of Speech, Language, and Hearing Research* 49.3, pp. 645–659.

Kent, Raymond D. and Houri K. Vorperian (2018). "Static measurements of vowel formant frequencies and bandwidths: A review". In: *Journal of Communication Disorders* 74, pp. 74–97.

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.

Lee, Jimin, Meghan Anne Littlejohn, and Zachary Simmons (2017). "Acoustic and tongue kinematic vowel space in speakers with and without dysarthria". In: *International Journal of Speech-Language Pathology* 19.2, pp. 195–204.

Mefferd, Antje S. and Jordan R. Green (2010). "Articulatory-to-Acoustic Relations in Response to Speaking Rate and Loudness Manipulations". In: *Journal of Speech, Language, and Hearing Research* 53.5, pp. 1206–1219.

Neel, Amy T. and Phyllis M. Palmer (2012). "Is Tongue Strength an Important Influence on Rate of Articulation in Diadochokinetic and Reading Tasks?" In: *Journal of Speech, Language, and Hearing Research* 55.1, pp. 235–246.

Opara, Józef, Andrzej Małecki, Elżbieta Małecka, and Teresa Socha (2017). "Motor assessment in Parkinson's disease". In: *Annals of Agricultural and Environmental Medicine* 24.3, pp. 411–415.

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/.

Skodda, Sabine, Wenke Visser, and Uwe Schlegel (2011). "Vowel Articulation in Parkinson's Disease". In: *Journal of Voice* 25.4, pp. 467–472.

Thompson, Austin and Yunjung Kim (2019). "Relation of second formant trajectories to tongue kinematics". In: *The Journal of the Acoustical Society of America* 145.4, EL323–EL328.

Tjaden, Kris, Jennifer Lam, and Greg Wilding (2013). "Vowel Acoustics in Parkinson's Disease and Multiple Sclerosis: Comparison of Clear, Loud, and Slow Speaking Conditions". In: *Journal of Speech, Language, and Hearing Research* 56.5, pp. 1485–1502.

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.

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, the Netherlands*
[2]*The Netherlands Cancer Institute, the Netherlands*
[3]*University of Amsterdam, the Netherlands*
t.b.tienkamp, t.rebernik, raoul.buurke, k.m.polsterer, m.b.wieling, d.abur[@rug.nl],
r.v.son@nki.nl, m.j.h.witjes, s.a.h.j.de.visscher[@umcg.nl]

Abstract

The impact of surgical treatment for tongue cancer is traditionally assessed with vowel formant metrics from read speech or sustained vowels. However, isolated speech might not fully reflect a speaker's typical speech. Here, we assessed the effect of speaking style (read vs. semi-spontaneous) on vowel acoustics of individuals pre- and post-surgery for tongue cancer. Eight individuals (3 females and 5 males) were recorded pre- and approximately six months post-surgery. We calculated the articulatory-acoustic vowel space (AAVS) during read speech (sentences) and semi-spontaneous speech (picture description). Results showed that the AAVS did not differ significantly pre- and post-surgery. Picture descriptions yielded a significantly smaller AAVS compared to the reading task, which was consistent pre- and post-surgery. Our findings suggest that both read and semi-spontaneous speech styles would be suitable to quantify the impact of surgical intervention for tongue cancer on vowel acoustics.

Keywords: speech production, vowel acoustics, tongue cancer

1. Introduction

Surgical intervention for tongue cancer often reduces tongue mobility (Lazarus et al. 2014; Tienkamp et al. 2024). Reduced tongue mobility may lead to more centralised speech where the acoustic distance between phonemes becomes smaller. Studies that assess the effect of surgery for tongue cancer on speech acoustics often use sustained vowels or isolated words and/or sentences over (semi-)spontaneous speech for their clinical feasibility and increased experimental control (Takatsu et al. 2017; Guo et al. 2023). Studies using isolated utterances have indicated that the vowel space area (VSA) is generally reduced in individuals with tongue cancer following surgical treatment (Balaguer et al. 2020; Takatsu et al. 2017; Guo et al. 2023). However, a recent study that used spontaneous speech did not find significant differences between the vowel formants of individuals treated for tongue cancer and control speakers (Tienkamp, van Son, and Halpern 2023). This raises the question to what extent the conflicting findings for VSA metrics in speakers treated for tongue cancer might result from differences in speaking style.

The choice of speech prompt (vowels/syllables or words) or speaking style (slow, read, or semi-spontaneous) has a considerable effect on the resulting speech output. For choice of speech prompt, individual syllables result in larger vowel spaces compared to words or sentences (van Son, Middag, and Demuynck 2018). For speaking style, larger vowel spaces are found when speakers are asked to read aloud a passage more slowly compared to their habitual speech rate (Turner, Tjaden, and Weismer 1995). In contrast, (semi-)spontaneous speech, which is primarily characterised by a faster speech rate, has resulted in the acoustic reduction of both vowels and consonants compared to read speech (Nakamura, Iwano, and Furui 2008; van Son and Pols 1999). Thus, while sustained vowels or read speech might allow for the recording of best-effort attempts as it elicits larger vowel spaces, more spontaneous speech better reflects daily conversational speech.

At present, no direct comparisons have been made between read and more spontaneous speech in speakers surgically-treated for tongue cancer. Yet, a better understanding of how speaking style affects vowel acoustics before and after surgery for tongue cancer can aid in the development of a standardised speech assessment protocol, which does not exist at present. Specifically, it is not clear which speaking style best captures changes in vowel acoustics following surgical treatment for tongue cancer.

The objective of this study was therefore to assess the effect of speaking style (read vs. semi-spontaneous) on the comprehensive acoustic vowel space in individuals undergoing surgical treatment for tongue cancer. To this end, we measured the articulatory-acoustic vowel space (AAVS) across sentence reading and across more spontaneously elicited speech (i.e., a picture description task) in individuals before and after surgery for tongue cancer. We predicted that the picture description task would yield a smaller AAVS (i.e., more reduced speech) compared to the sentence reading task. Moreover, we predicted an overall reduction of the AAVS following treatment as compared to pre-treatment due to a surgery-induced reduction in tongue mobility. Due to a lack of prior studies on the topic, we did not formulate any specific predictions regarding the interaction between speaking style and treatment.

2. Methods

2.1. Participants

The present study is part of a larger project approved by the institution's Medical Ethics Review Board (NL79242.042.21). All participants provided written informed consent before their participation. Eight native speakers of Dutch (five males, three

females) with a mean age of 62.1 years (range: 41-77) completed data collection both pre- and post-surgery and were included in this study. Participants were tested a few days before and approximately six months after surgical treatment. Speakers were treated for T1 (n=5), T2 (n=2) or T3 (n=1) tongue tumours located either on the mid-line of the tongue (S07) or the lateral side of the tongue (all other speakers). T-stages can range from T1 (smallest) to T4 (largest). For six speakers, the tumour was localised on the anterior $2/3$ of the tongue, whereas for two speakers (S02 and S04), the tumour was localised on the posterior $1/3$ of the tongue. The tongue was reconstructed using a radial forearm free flap for two speakers (S01 and S02), whereas the wound was locally closed for other speakers. One speaker received (chemo)radiation post-surgery (S02) and was recorded six months after the last radiation session to ensure a comparable time post-treatment. Table 1 shows the demographic and clinical information of all speakers.

Table 1: *Speaker demographics and clinical information. F = female, M = male, Anterior = anterior* $2/3$ *of the tongue. Posterior = posterior* $1/3$ *of the tongue.*

Speaker	Sex	Age	T-stage	Location
S01	F	75	T3	Anterior
S02	M	41	T2	Posterior
S03	M	54	T1	Anterior
S04	F	77	T1	Posterior
S05	M	55	T1	Anterior
S06	M	68	T2	Anterior
S07	F	61	T1	Anterior
S08	M	62	T1	Anterior

2.2. Procedures

All speakers were recorded in the mobile sound booth SPRAAKLAB (Wieling, Rebernik, and Jacobi 2023) and were fitted with an omni-directional microphone (Shure MX-153T) angled 45° from the mouth with a seven centimetre mic-to-mouth distance. Their speech was recorded at a 22,050 Hz sampling rate. To elicit semi-spontaneous speech, participants were asked to describe two pictures in detail using their habitual speaking style: the Cookie Theft picture (Goodglass, Kaplan, and Weintraub 2001) and the Cat Rescue picture (Nicholas and Brookshire 1993). To elicit read speech, participants were asked to read aloud 15 phonemically-balanced sentences with the phonemes of Dutch at the frequency the phonemes typically occur (Luts et al. 2014). In the case of a misreading, speakers were asked to repeat the sentence and only the correctly read instance was used for analysis.

2.3. Acoustic analysis

We used the articulatory-acoustic vowel space (AAVS) in this study (Whitfield and Goberman 2014) to quantify vowel articulation in each speaking style. An advantage of the AAVS over point-based metrics, such as the VSA, is that the AAVS can be computed over full trajectories of running speech (e.g., picture descriptions and full sentences). For this reason, the AAVS takes all vowels into account, thus increasing ecological validity. We calculated the AAVS according to methods established in prior work and developed a custom semi-automatic pipeline (Whitfield and Goberman 2014; Abur, Perkell, and Stepp 2022). First, all instances of 'uh' and 'uhm' were manually removed from the picture description recordings. Next, we removed all voiceless segments using a custom script in Praat (version 6.3.1; Boersma and Weenink 2023). Continuous first and second formant frequency (F_1 and F_2) traces were extracted in Praat from the voiced segments using a script based on Carignan (2022). As formant frequency tracking accuracy is considerably influenced by both speaker and vowel characteristics, the Carignan (2022) script extracts the 'optimal' formant frequency by calculating the F_1 and F_2 using formant ceilings ranging from 3500-6000 Hz with 50 Hz intervals (see e.g., Escudero et al. 2009), time steps of 5 ms, and 25 ms time windows. From these 51 possible formant values (one associated with each ceiling), those two standard deviations away from each mean formant value were removed. From the remaining formant frequencies, the median value was taken as the optimal formant frequency for each given 5 ms time step (representing a single data point).

The resulting formant trajectories were filtered using a median absolute deviation filter, removing data points 2.5 times away from the median absolute deviation of the dataset (5,584 rows, 1.8%).[1] We calculated the AAVS on a mel-scale for each speaker at each assessment point, and for each speaking style (four AAVS values per speaker). The formant trajectories of both picture descriptions were combined to calculate one AAVS value. The AAVS was calculated as the square-root of the product of the squared variance of the formant tracks and the unshared variance between them (see equation (1)). The unshared variance was calculated by subtracting the R^2 of a linear model with F_1 predicting F_2 from 1.

$$\text{AAVS} = \sqrt{(\sigma_{F_1})^2 \times (\sigma_{F_2})^2 \times (1 - \text{R}^2)} \quad (1)$$

2.4. Statistical analysis

The data were analysed using linear mixed-effects regression in R (version 4.3.2; R Core Team 2023; Bates et al. 2015; Kuznetsova, Brockhoff, and Christensen 2017). Our hypothesis-testing model included the AAVS as a function of surgery (pre vs. post surgery) in interaction with style (read speech vs. semi-spontaneous), together with a by-speaker random intercept. We further assessed the influence of speaker sex and articulation rate (number of syllables / phonation time in seconds) in an exploratory modeling procedure, as these variables can impact vowel acoustics. The articulation rate was calculated using a Praat script by De Jong and Wempe (2009). All numerical variables were centered around the mean and the α-level was set at 0.05. We concluded our analysis by verifying the model's assumptions and employing model criticism (Fox and Weisberg 2019). Data points with an absolute residual exceeding 2.5 standard deviations from their fitted value were removed. We only used this trimmed dataset when outliers drove the absence or presence of statistically significant effects (Baayen 2008).

3. Results

Our results are based on the trimmed dataset that removed one data point from the analysis (3%). An overview of the AAVS values per style and time point is provided in Figure 1-A. The AAVS post-treatment was not significantly smaller compared to

[1] Additional manual filtering only removed an extra 600 rows (0.2%). The correlation between the AAVS with and without manual filtering was $r = .99$ and our subsequent results were nearly identical. We therefore use the AAVS values without manual filtering.

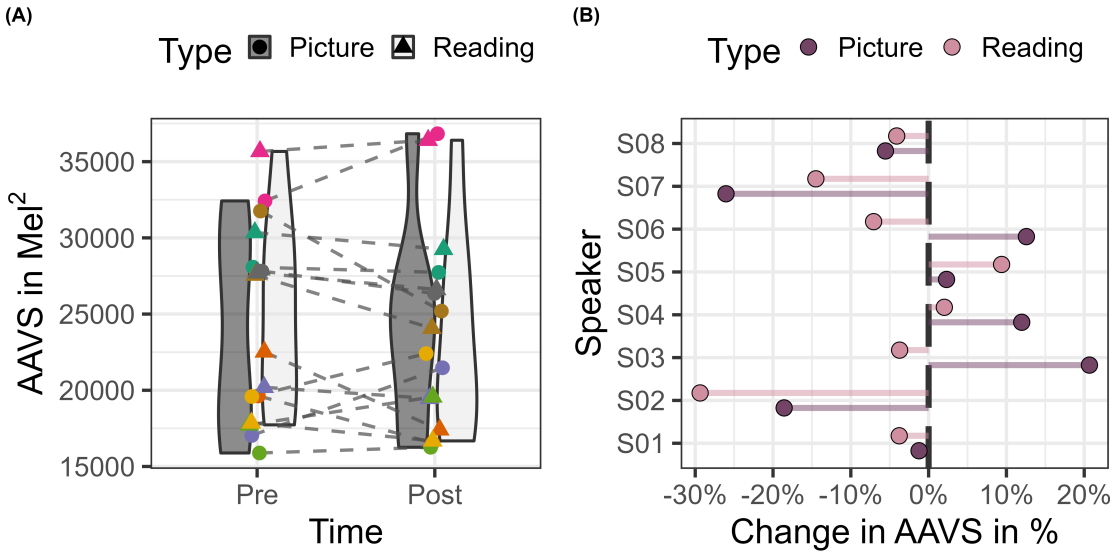


Figure 1: **(A)** Articulatory-acoustic vowel space (AAVS) per time point and speaking style. Different colours represent individual speakers, different shapes represent the speaking styles (circles = picture description, triangles = sentence reading). **(B)** Change in AAVS in percentage compared to pre-treatment between both speaking styles per speaker. A negative value indicates that the AAVS was smaller post-treatment compared to pre-treatment. A positive value indicates an increase in AAVS.

pre-treatment ($p = .66$). On average, semi-spontaneous speech yielded a significantly smaller AAVS compared to read speech (β = -2,621 mel^2, T = -2.7, CI = [-4,471, -529], $p = .02$). There was no significant interaction between time and style ($p = .22$). Figure 1-B shows the change in AAVS in percentage compared to pre-treatment per speaking style and speaker. Decreases in AAVS for both speaking styles were observed for four speakers (S01, S02, S07, S08) post-surgery, with the largest decrease in AAVS for speakers S02 and S07.

Our exploratory analysis revealed a significant effect of sex which indicated that, on average, males had a lower AAVS compared to females (β = -11,026 mel^2, T = -3.9, CI = [-16,793, -5,250], $p < .01$). A significant effect of articulation rate indicated a positive relationship between articulation rate and AAVS (β = 5,184 mel^2, T = 2.3, CI = [870, 9,633], $p < .05$). With the inclusion of the exploratory variables, the fixed effect of speaking style became significant.

4. Discussion and conclusion

We assessed the effect of speaking style on the articulatory-acoustic vowel space (AAVS) of individuals with tongue cancer pre- and post-surgical intervention. The results suggest that the surgical intervention did not impact the overall vowel space for the speakers included in this study. This is not in line with previous work that reported a reduced VSA following treatment compared to pre-treatment for tongue cancer (Guo et al. 2023; Takatsu et al. 2017). One important difference, compared to earlier work, is that the speakers included in our study were mostly treated for smaller tumours (T1) whereas the studies by Guo et al. (2023) and Takatsu et al. (2017) also included individuals with large tumours (T4). To rule out the possibility of pre-treatment speech impairments influencing our findings, we verified that our speakers had typical AAVS values before treatment by comparing them to those of Dutch typical speakers (Hoekzema et al. 2024, current proceedings).

The absence of a reduced AAVS could stem from varying post-treatment changes among speakers, as some had an increase in AAVS following treatment (e.g., S03, S04, S05) whereas others showed a decrease (e.g., S02 and S07). The largest increases post-surgery were seen for the semi-spontaneous speech style in speakers treated for anterior tumours. The surgery may have relieved tumor-related discomfort without significantly affecting articulatory function, potentially resulting in increased range of motion during speech post-treatment. In contrast, the two speakers with the largest decrease in AAVS post-surgery (S02 and S07) were treated for either a posterior tumour or a tumour located on the mid-line of the tongue, which seem to have a more pronounced effect on vowel articulation.

On average, speakers with faster articulation rates had a larger AAVS which might seem contradictory at first, as a faster articulation rate typically results in a smaller VSA (Turner, Tjaden, and Weismer 1995). However, speakers whose speech was more affected by surgery might have slowed their speech rate as a compensatory strategy, whereas speakers whose speech was less affected may have remained at their habitual articulation rate and preserved the size of their acoustic working space.

The results of our study further suggest that the AAVS can capture differences induced by speaking style. Previous work showed that the AAVS yielded larger AAVS values in clear speech compared to typical speech during a reading passage (Whitfield and Goberman 2014). We extend these findings by showing that spontaneously elicited speech from picture descriptions resulted in a smaller AAVS compared to a reading task with individual sentences. While it is still possible that speakers produced 'clear' speech during the picture description, 'clear' semi-spontaneous speech still elicits smaller formant ranges compared to 'clear' read speech (Hazan and Baker 2010). The effect of speaking style on AAVS did not change as

a result of surgery for the speakers in our study.

It should be noted that our results are based on a small group-level assessment, which is a limitation of our study. A second limitation is that the phonemic content of both speaking styles was not identical. However, we tried to control for this by including sentences that included a distribution of Dutch phonemes at the frequency the phonemes typically occur.

To conclude, to quantify the effect of surgical treatment for tongue cancer on the acoustic vowel space, our results suggest that both reading and semi-spontaneous speech styles would be suitable prompts to use.

5. Acknowledgements

The authors would like to express their gratitude to all speakers who were willing to participate in this study. This work was supported by a University of Groningen CLCG PhD grant awarded to the first author and by the university's Research School of Behavioral and Cognitive Neurosciences. This work was also supported by a research grant from Atos Medical (Hörby, Sweden) awarded to the department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute.

6. References

Abur, Defne, Joseph S. Perkell, and Cara E. Stepp (2022). "Impact of Vocal Effort on Respiratory and Articulatory Kinematics". In: *Journal of Speech, Language, and Hearing Research* 65.1, pp. 5–21.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Balaguer, Mathieu, Timothy Pommée, Jérôme Farinas, Julien Pinquier, Virginie Woisard, and Renée Speyer (2020). "Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: Systematic review". In: *Head & Neck* 42.1, pp. 111–130.

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 [computer programme]*. Version Number: 6.3.1. URL: https://praat.org.

Carignan, Christopher (2022). *Formant Optimization*. URL: https://github.com/ChristopherCarignan/formant-optimization.

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.

Escudero, Paola, Paul Boersma, Andréia Schurt Rauber, and Ricardo A. H. Bion (2009). "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese". In: *The Journal of the Acoustical Society of America* 126.3, pp. 1379–1393.

Fox, John and Sanford Weisberg (2019). *An R Companion to Applied Regression*. Third edition. Thousand Oaks CA: Sage.

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

Guo, Kaixin, Yudong Xiao, Wei Deng, Guiyi Zhao, Jie Zhang, Yujie Liang, Le Yang, and Guiqing Liao (2023). "Speech disorders in patients with Tongue squamous cell carcinoma: A longitudinal observational study based on a questionnaire and acoustic analysis". en. In: *BMC Oral Health* 23.1, p. 192.

Hazan, Valerie and Rachel Baker (2010). "Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style?" In: *DiSS-LPSS Joint Workshop 2010*, pp. 7–10.

Hoekzema, Nikki, Teja Rebernik, Thomas B. Tienkamp, Sasha Chaboksavar, Valentina Ciot, Annetje Gleichman, Roel Jonkers, Aude Noiray, Martijn B. Wieling, and Defne Abur (2024). "Assessing differences in articulatory-acoustic vowel space in Parkinson's disease by sex and phenotype". In: *Proceedings of the 13th International Seminar on Speech Production*.

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.

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.

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.

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/.

Takatsu, Jun, Nobuhiro Hanai, Hidenori Suzuki, Masahiro Yoshida, Yasuhiro Tanaka, Seiya Tanaka, Yasuhisa Hasegawa, and Masahiko Yamamoto (2017). "Phonologic and Acoustic Analysis of Speech Following Glossectomy and the Effect of Rehabilitation on Speech Outcomes". In: *Journal of Oral and Maxillofacial Surgery* 75.7, pp. 1530–1541.

Tienkamp, Thomas, Teja Rebernik, Bence Halpern, Rob van Son, Martijn Wieling, Max Witjes, Sebastiaan de Visscher, and Defne Abur (2024). "Quantifying Changes in Articulatory Working Space in Individuals Surgically Treated for Oral Cancer with Electromagnetic Articulography". In: *Journal of Speech, Language and Hearing Research* 67.2, pp. 384–399.

Tienkamp, Thomas, Rob van Son, and Bence Halpern (2023). "Objective speech outcomes after surgical treatment for oral cancer: An acoustic analysis of a spontaneous speech corpus containing 32.850 tokens". In: *Journal of Communication Disorders* 101, p. 106292.

Turner, Greg S., Kris Tjaden, and Gary Weismer (1995). "The Influence of Speaking Rate on Vowel Space and Speech Intelligibility for Individuals With Amyotrophic Lateral Sclerosis". In: *Journal of Speech, Language, and Hearing Research* 38.5, pp. 1001–1013.

van Son, Rob, Catherine Middag, and Kris Demuynck (2018). "Vowel Space as a Tool to Evaluate Articulation Problems". In: *Proceedings of Interspeech 2018*. ISCA, pp. 357–361.

van Son, Rob and Louis C. W. Pols (1999). "An acoustic description of consonant reduction". In: *Speech Communication* 28.2, pp. 125–140.

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.

Wieling, Martijn, Teja Rebernik, and Jidde Jacobi (2023). "SPRAAKLAB: a mobile laboratory for collecting speech production data". In: *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023)*, pp. 2060–2064.

Assessing Differences in Articulatory-Acoustic Vowel Space in Parkinson's Disease by Sex and Phenotype

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 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, t.b.tienkamp@rug.nl, s.a.chaboksavar@student.rug.nl, valentinaciot@gmail.com, annetje.gleichmann@gmail.com, r.jonkers@rug.nl, aude.noiray@univ-grenoble-alpes.fr, m.b.wieling@rug.nl, d.abur@rug.nl

Abstract

The goal of this study was to determine whether articulatory-acoustics differ between individuals in the tremor-dominant (TD) and postural instability/gait difficulty (PIGD) phenotypes of Parkinson's disease (PD). The study included 31 individuals with PD (21 TD, 10 PIGD) and 29 control speakers (CS) who were all Dutch native speakers. A read speech task and a semi-spontaneous speech task were completed, and the Articulatory-Acoustic Vowel Space (AAVS) was calculated for both tasks. Results showed no significant difference in AAVS between the overall control group and PD for either phenotype. Follow-up analyses, pooling speech data from our prior study (+27 PD, +23 CS), demonstrated a significantly lower AAVS in males with PD compared to controls and no group differences for females. Thus, articulatory-acoustic changes may be more pronounced for male compared to female speakers with PD, but may not differ by PD phenotype.

Keywords: Parkinson's Disease, Phenotype, Speech Acoustics, Articulatory-Acoustic Vowel Space

Introduction

Parkinson Disease (PD) is a neurodegenerative disorder that is associated with a degeneration of dopaminergic neurons (Tysnes & Storstein, 2017). PD is a multisystem disorder, characterized by both motor impairments, such as muscle rigidity, tremor, and slowness of movement, as well as non-motor impairments, such as cognitive impairments and fatigue (Jankovic, 2008). The symptoms and progression of the disease vary depending on the individual, with various factors, such as sex (Iwaki et al., 2021), age (Wickremaratchi et al., 2009), and cognition (Sollinger et al., 2010) playing a role.

Due to the differing symptomatologies, different distinct clinical phenotypes of PD have been identified, with a frequent distinction being made between Tremor Dominant (TD) versus Postural Instability/Gait Difficulty (PIGD) phenotypes of PD (Stebbins et al., 2013). The TD phenotype is primarily characterized by the presence of tremor in the limbs, while the PIGD phenotype is primarily characterized by gait disturbance, postural instability, and rigidity (ibid.).

A common problem faced by most individuals with PD (IwPD), regardless of phenotype, are speech impairments, including respiration, laryngeal impairments, and articulation (see also Pinto et al., 2004, Broadfoot et al., 2019). At the level of articulatory impairments, IwPD are often impaired in their vowel articulation, which has been shown to be potentially reduced when measured with acoustic measures such as the Vowel Space Area (VSA), Vowel Articulation Index (VAI; Sapir et al., 2011) and Articulatory-Acoustic Vowel Space (AAVS; Whitfield & Goberman, 2014). While many studies have found a smaller VSA in IwPD compared to control speakers (Tjaden et al., 2013; Skodda et al., 2011; Leung et al., 2018), some studies have shown no differences in VSA between the two groups (e.g., Douadi et al., 2022). However, as the VSA is sensitive to interspeaker variability (Sapir et al., 2011), other studies have used new vowel formant metrics that would be more likely to capture minute group differences in vowel production. One of these metrics, the VAI, is a measure for vowel centralization that is less sensitive to interspeaker differences and has been shown to be smaller in IwPD compared to control speakers (Sapir et al., 2011; Skodda et al., 2011).

However, both VSA and VAI rely on having clearly elicited and segmented vowels, even though IwPD potentially experience more issues in spontaneous speech tasks than in read speech (e.g., Rusz et al., 2013). It is therefore crucial to assess sentence-level speech metrics when investigating speech in IwPD. The Articulatory-Acoustic Vowel Space (AAVS), introduced as a measure of an individual's working formant space (Whitfield & Goberman, 2014) is a vowel space metric that is sensitive to differences between groups, calculated at a sentence-level and is not point-based (Whitfield, 2019). In prior work, IwPD showed significantly smaller AAVS compared to control speakers in one study (Whitfield & Goberman, 2014; based on a sample of 12 IwPD and 10 CS) but another study found no group differences in AAVS (Houle et al., 2023; based on a sample of 68 IwPD and 68 CS).

A potential explanatory variable for the conflicting results on AAVS findings in IwPD is the IwPD phenotype, which has not been previously considered in AAVS studies in IwPD. Prior work has suggested more severe speech impairments in PIGD than TD phenotypes of PD when compared to control speakers. Specifically, one study found slower speaking rates during a monologue in PIGD compared to TD speakers with PD (Tykalová et al., 2020), while other work found a faster DDK rate (in syllables/s) in PIGD compared to TD (Rusz et al., 2023). Another study, using VSA and VAI based on corner vowels extracted from a reading passage, suggested a negative correlation between VSA and VAI, and high bradykinesia and rigidity subscores, but no significant correlation between PIGD or tremor subscores and the VSA (Skrabal et al., 2022). However, no study to-date has

assessed sentence-level vowel metrics in PD compared to controls while considering PD phenotypes. Assessing sentence-level articulatory differences allows us to analyze speech across a wider range of vowel productions and is more ecologically valid than using vowels in isolation.

The current study therefore assessed whether there is a difference in sentence-level vowel production between PD phenotypes, as well as compared to control speakers, as quantified via the AAVS (Whitfield & Goberman, 2014). In addition, we assessed whether other variables, including task (reading vs. semi-spontaneous 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 smaller AAVS) than control speakers (CS), but a comparable AAVS between the TD phenotype and control speakers. We additionally expected a larger AAVS in female than male speakers, regardless of group (Whitfield & Goberman, 2014; Houle et al., 2023).

Methods

The present study forms part of a larger study, approved by our institutional Medical Ethics Review Board (NL72589.042.21).

Participants

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). To ensure the participants' ability to give consent, only individuals with a MoCA score of 22 or higher were included in the study (Karlawish et al., 2013).

Participants underwent an age-appropriate pure tone hearing screening at 25dB for tones at or below 1000 Hz, and 40 dB for tones at 2000 Hz and above (Schow, 1991). This screening was conducted without hearing aids. We subsequently 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). Where applicable, speech tasks were completed while the participants wore their hearing aids and therefore had corrected-to-normal hearing (hearing aids worn by 3 CS, 4 IwPD). Table 1 summarizes participant demographics.

Table 1: *Participant demographics, separated by group (PIGD: postural instability/gait difficulty, TD: tremor-dominant, CS: control speakers). Sex: M (male), F (female). Hearing: NtM (None to Mild), MtS (Moderate to Severe). MoCA scores: maximum 30 points (22–25 points: potential Mild Cognitive Impairment (MCI), 26—30 points: no Mild Cognitive Impairment (nMCI)).*

Variable	PIGD	TD	CS
Sex	7 M, 3 F	11 M, 10 F	15 M, 14 F
Age (years)	67.8 ± 8.3	73.1 ± 4.7	68.1 ± 7.3
Hearing	4 NtM 6 MtS	9 NtM 12 MtS	9 NtM 20 MtS
MoCA	4 MCI 6 nMCI	8 MCI 13 nMCI	7 MCI 22 nMCI
MDS-UPDRSIII	21-71 pt.	11-61 pt.	-

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). This allowed us to assess the participants' motor symptom severity (part III of the scale) as well as classify the motor phenotype. Following Stebbins and colleagues (2013), our sample included 22 TD (11 male, 10 female; MDS-UPDRS part III range: 11-61 points) and 10 PIGD (7 male, 3 female; MDS-UPDRS part III: 21-71 points) IwPD. All IwPD completed the experimental tasks while ON levodopa.

Procedure

The study took place in two sessions; the data reported in this paper was collected at the beginning of the second session. The participants were seated in the sound-dampened booth of SPRAAKLAB, the mobile laboratory of the Faculty of Arts, University of Groningen (Wieling et al., 2023). After placing a Shure MX153 earset microphone seven centimetres away from the participant's mouth, they were asked to read the Dutch version of the North Wind and the Sun passage (Roach, 2004), to describe the Cookie Theft picture (Goodglass & Kaplan, 1983), and to answer four questions eliciting spontaneous speech (not reported in this paper). The acoustic data was recorded in Praat v6.2.18 (Boersma & Weenink, 2022). Data was collected with a sampling rate of 44.1 kHz and digitized via Focusrite Scarlett Solo (2nd gen).

Data pre-processing

The AAVS measures a speaker's vowel production based on continuously sampled formant trajectories in running speech. The recordings were first cut to remove any speech resulting from experimenter instructions, followed by a removal of all pauses and voiceless segments using a customized Praat script. We extracted formants automatically using a Praat script that determines speaker-specific and segment-specific optimal ceiling levels using the Burg algorithm, with five millisecond timesteps in a 25 ms time window (Carignan, 2022). AAVS was subsequently calculated in mels for two tasks, namely the North Wind and the Sun passage ('read speech') and the Cookie Theft picture description ('semi-spontaneous speech'), following the methods and formulas as specified in Whitfield and Goberman (2014) and Abur et al. (2022).

Statistical Analysis

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 PD, PIGD PD, CS) as the main fixed effect variable, and sex as an additional fixed effect variable. We included a by-participant random intercept. In our exploratory analysis, we further assessed the effect of age, task (read vs. semi-spontaneous 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 difference in AAVS between control speakers and the PIGD (β = -3,858 mel^2, t = -1.9, p = 0.06,

Cohen's d = -0.5) or TD (β = -1,036, t = -0.7, p = 0.5, Cohen's d = -0.2) phenotype groups.

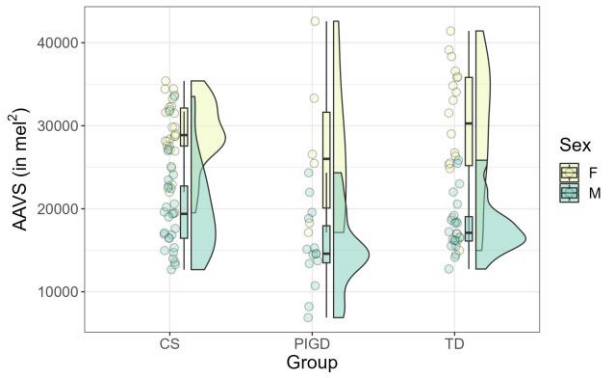


Figure 1: *Difference in AAVS (in mel²) depending on phenotype (CS, PIGD, TD) and sex (male (M), female (F)).*

There was a significant effect of sex on AAVS overall, with females having a significantly larger AAVS than males (β = 10,763 mel², t = 7.5, p < 0.001, Cohen's d = 2.0). The interaction between sex and group did not significantly improve the model, however (p = 0.7). 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. There was therefore no significant effect of age (p = 0.99), cognition (p = 0.4), task choice (p = 0.8), or hearing impairment (p = 0.4) on AAVS observed.

To test whether there was an overall difference between control speakers and IwPD, we ran an additional model with a binary distinction between the CS and (combined) PD groups. This model, likewise, did not show a significant effect of group on AAVS (β = -1,932 mel², t = -1.37, p = 0.17, Cohen's d = -0.36).

Exploratory analysis of sex

Our results conflicted with those of Tienkamp and colleagues (2024, current volume), as they found a significantly smaller AAVS in IwPD than CS. However, as the data used in the paper by Tienkamp and colleagues (2024) stems from the same lab, using the same reading task (i.e., The North Wind and the Sun) but different participant groups, we had the unique opportunity to conduct an additional analysis.

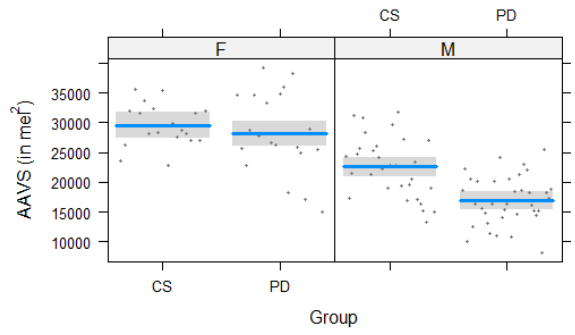


Figure 2: *Difference in AAVS (in mel²) depending on group (CS, PD) and sex (male (M), female (F)).*

We pooled the datasets in order to strengthen the power of the current investigation. The joint analysis therefore included 58 IwPD (21 female, 37 male) and 52 CS (19 female, 33 male). A linear model, assessing the effect of the interaction between group and sex on AAVS in mel², showed that male IwPD had a smaller AAVS compared to male CS (β = -4278 mel², t = -2.2, p = 0.03) while female IwPD and female CS had a comparable AAVS (β = -1345 mel², t = -0.9, p = 0.38). Figure 2 shows the effect of group and sex on AAVS. Unfortunately, we have no disease severity measurement or phenotype indication for the dataset used by Tienkamp and colleagues (2024); thus, it is not clear if there is an impact of PD phenotype on these results.

Discussion

The purpose of this study was to examine whether vowel articulation is differentially impacted in PD by an individual's clinical phenotype (TD or PIGD) compared to controls. Our study results indicate no significant impact of PD phenotype on the AAVS: while there was a trend towards the PIGD phenotype having lower AAVS than the TD phenotype or control speakers, the number of speakers in the PIGD group was too small and contained too many male speakers (7M, 3F) to draw reliable conclusions.

We likewise did not find any differences between CS and IwPD when the phenotypes were grouped. This finding aligns with the results of Houle and colleagues (2023), but conflicts with those of Whitfield and Goberman (2014) and Tienkamp and colleagues (2024, current proceedings), who report a smaller AAVS in IwPD compared to CS.

However, a linear model assessing the effect of the interaction between group and sex on AAVS, using pooled data from a study with the same methods and different speakers with PD (Tienkamp et al., 2024), revealed that male IwPD had a smaller AAVS compared to male CS, while female IwPD and female CS had a comparable AAVS. As we do not have motor severity scores for the entire dataset, it remains unclear whether our current finding indicates that more articulation impairments are actually present in male than female IwPD, or that our sample included more severely motor impaired male IwPD than female IwPD. This is not the first time a potential difference was shown in the articulation of male and female IwPD, however, as a study by Skodda and colleagues (2011) previously showed that only male IwPD showed a smaller VSA compared to CS, while both female and male IwPD showed smaller VAI values compared to CS.

Overall, following previous studies (Whitfield & Goberman, 2014; Houle et al., 2023), female speakers exhibited a significantly larger AAVS than male speakers. However, our current study did not find an effect of any other factors, such as cognition, hearing impairment, age or task choice on the AAVS. The latter finding, especially, is informative for future studies investigating articulation in IwPD. While prior studies used the Rainbow Passage reading task to assess the AAVS, our study also included a more ecologically valid semi-spontaneous speech task next to a read speech task. As the two tasks were comparable in terms of the AAVS, this indicates that choosing a semi-spontaneous speech task is a suitable choice for researchers who wish to evaluate differences in the articulatory-acoustic vowel space as part of a larger battery evaluating multiple subsystems. Alternatively, those wishing to conduct detailed acoustic analyses can use a reading task with a comparable text across participants.

A limitation of our study is the unbalanced participant sample, with a relatively small PIGD group (10 participants) compared to the TD group (21 participants) and the control group (29 participants), thereby limiting the generalizability of our findings.

Conclusion

The current study provided a look into the understudied sentence-level vowel production in PD phenotypes and control speakers, using the Articulatory Acoustic Vowel Space (AAVS) measure. While the results remain inconclusive and show no significant differences between PD phenotypes (TD or PIGD) and CS groups, they provide a first glimpse into sentence-level articulation of speakers of different IwPD phenotypes and underscore the importance of keeping sex and phenotype in mind when assessing speech motor control in IwPD.

References

Abur, D., Perkell, J. S., & Stepp, C. E. (2022). Impact of Vocal Effort on Respiratory and Articulatory Kinematics. *Journal of Speech, Language, and Hearing Research, 65*(1), 5–21.

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

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

Broadfoot, C. K., Abur, D., Hoffmeister, J. D., Stepp, C. E., & Ciucci, M. R. (2019). Research-based Updates in Swallowing and Communication Dysfunction in Parkinson Disease: Implications for Evaluation and Management. *Perspectives of the ASHA special interest groups*, 4(5), 825–841.

Carignan, C. (2022). Formant Optimization. GitHub repository. Retrieved from: https://github.com/ChristopherCarignan/formant-optimization.

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).

Goodglass, H., & Kaplan, E. (1983). The Assessment of Aphasia and Related Disorders, *Lea and Febiger*, ed. 2, Philadelphia, PA.

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.

Iwaki, H., Blauwendraat, C., Leonard, H. L., Makarious, M. B., Kim, J. J., Liu, … Nalls, M. A. (2021). Differences in the Presentation and Progression of Parkinson's Disease by Sex. *Movement disorders : official journal of the Movement Disorder Society, 36*(1), 106–117.

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

Karlawish, J., Cary, M., Moelter, S. T., Siderowf, A., Sullo, E., Xie, S., & Weintraub, D. (2013). Cognitive impairment and PD patients' capacity to consent to research. *Neurology, 81*(9), 801–807.

Leung, N., Tong, E., Ng, M. (2018) Vowel characteristics associated with Parkinson's disease in Cantonese. *Movevement Disorders, 33*(2).

Logemann, J. A., Fisher, H. B., Boshes, B., & Blonsky, E. R. (1978). Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *The Journal of Speech and Hearing Disorders*, 43(1), 47–57.

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.

Pinto, S., Ozsancak, C., Tripoliti, E., Thobois, S., Limousin-Dowsey, P., & Auzou, P. (2004). Treatments for dysarthria in Parkinson's disease. *The Lancet. Neurology, 3*(9), 547–556.

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

Roach, P. (2004). British English: Received Pronunciation. *Journal of the International Phonetic Association, 34*(2), 239-245.

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.

Sapir, S., Ramig, L., Spielman, J., & Fox, C. (2011). Acoustic metrics of vowel articulation in Parkinson's disease: vowel space area (VSA) vs. vowel articulation index (VAI). In *Proceedings of the 7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications.*

Schow, R. L. (1991). Considerations in Selecting and Validating an Adult/Elderly Hearing Screening Protocol. *Ear and Hearing, 12*(5), 337-448.

Skodda, S., Visser, W., & Schlegel, U. (2011). Vowel articulation in Parkinson's disease. *Journal of Voice, 25*(4), 467–472.

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*, *8*, 137.

Sollinger A.B., Goldstein F.C., Lah J.J., Levey A.I., & Factor S.A. (2010). Mild cognitive impairment in Parkinson's disease: subtypes and motor characteristics. *Parkinsonism and Related Disorders, 16*(3), 177-80.

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.

Tienkamp, T., Rebernik, T., Jacobi, J., Wieling, M., & Abur, D. (2024). The impact of electromagnetic articulography sensors on the articulatory-acoustic vowel space in speakers with and without Parkinson's Disease. In *Proceedings of the 13th International Seminar on Speech Production.*

Tjaden, K., & Wilding, G. (2011). Speech and pause characteristics associated with voluntary rate reduction in Parkinson's disease and Multiple Sclerosis. *Journal of communication disorders, 44*(6), 655–665.

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.

Tysnes, O.B., Storstein, A. (2017). Epidemiology of Parkinson's disease. *Journal of Neural Transmission, 124*(8), 901-905.

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.

Whitfield, J. A. (2019). Exploration of metrics for quantifying formant space: implications for clinical assessment of Parkinson's disease. *Perspectives of the ASHA Interest Groups.*

Wickremaratchi, M. M., Perera, D., O'Loghlen, C., Sastry, D., Morgan, E., Jones, A., Edwards, P., Robertson, N. P., Butler, C., Morris, H. R., & Ben-Shlomo, Y. (2009). Prevalence and age of onset of Parkinson's disease in Cardiff: a community based cross sectional study and meta-analysis. *Journal of neurology, neurosurgery, and psychiatry, 80*(7), 805–807.

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

Abstract

This study presents an innovative approach to speech breathing analysis, emphasizing the potential of Electromagnetic Articulography (EMA) as a viable tool. We compared the widely used Respiratory Inductive Plethysmography (RIP) with EMA by collecting speech breathing data from 18 speakers during sustained vowel productions of /a/ under habitual and loud speech conditions. Our findings indicate that EMA signals can effectively track temporal patterns of speech breathing movements, which do not differ from the RIP system. With this study, we would like to emphasize the potential of using (existing) EMA systems in laboratories to analyze speech breathing patterns. This paper explores the advantages and opportunities that arise from integrating EMA systems into speech breathing research. The findings suggest that such integration can enhance our understanding of speech production and contribute to advancements in related fields.

Keywords*: speech production, speech breathing, inductive plethysmography, electromagnetic articulography*

1. Introduction

The respiratory inductive plethysmography (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 artifacts 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, which has been recently developed by Columbi Computers AB (Sweden) for the RespTrack system. To simultaneously capture kinematic speech data, one is currently dependent on using two systems, such as RIP and e.g., an Electromagnetic Articulograph (EMA) as it has been done by e.g., Rasskazova et al. (2019).

Here, 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 artifacts. We conducted a study comparing the RIP system (Inductotrace®) and the EMA system (Carstens AG501) to track and analyze speech breathing patterns. The goal was to assess the similarity of the kinematic trajectories for capturing speech breathing patterns recorded by both systems. The data used for comparisons are sustained vowel productions in two different conditions, i.e., in habitual and loud speech.

In a first step, we analyze data from all applied EMA sensors to identify the most suitable sensors for accurately tracking speech breathing. This initial assessment ensures that the selected sensors provide reliable and precise measurements. The second step involves comparing the signals obtained from both the RIP system and the EMA system. By examining the signals from these two systems, we evaluated the consistency and accuracy of the EMA system in capturing speech breathing patterns. Finally, in the third step, we identify similarities in the signals to analyze the robustness of the tracking methods.

By conducting this comprehensive analysis, we aim to highlight the reliability and effectiveness of the EMA system for tracking speech breathing. The findings from this study will contribute to advancing research in speech production and to enhance our understanding of the intricate mechanisms involved in speech breathing.

2. Methods

2.1. Participants

We collected acoustic and kinematic data from 18 native German speaking participants (9 males, 9 females). The age ranged from 23 to 54 years with a mean age of 33 years.

2.2. Experimental Set-up

The kinematic breathing data were collected using the (a) EMA (AG 501) and (b) RIP (Inductotrace®) at the same time with a sampling rate of 1250 Hz. To track breathing data with EMA, sensors were placed at different positions and fixed with tape (**Figure 1**). One sensor on the lowest vertebra of the cervical spine functioned as the reference sensor. Sensors on the sternum and three on the chest were used to track (speech) breathing kinematics. Sensors tracking thorax movements were positioned at the axilla level on the chest (on clothes); one in the middle and two at the height of each papilla. After placing the EMA sensors (**Figure 1** left), the RIP band (only upper band for thorax movement) was put around the participants' chest (**Figure 1** right). Three different band sizes were used (7 x small, 5 x medium, and 6 x regular), thus representing different body sizes.

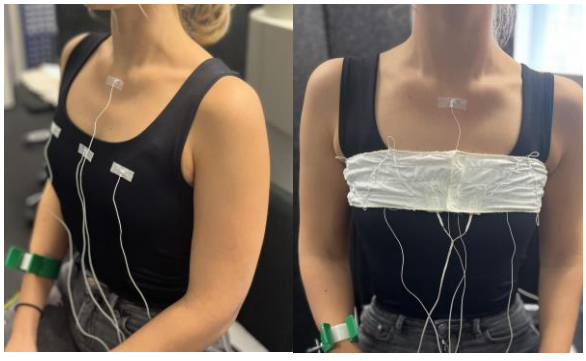


Figure 1: *EMA sensors on subject – (left) before the RIP belt is put on and (right) with the RIP belt put on.*

2.3. Speech Material

In this paper, only data of sustained productions of the vowel /a/ in habitual and loud speech are presented. The data analyzed here is part of a larger data set. Participants were asked to take a deep breath and to produce maximum phonation of the vowel /a/ in habitual speech and loud speech. Tracking of speech loudness was done via a sound level meter that was set up 1.25m away from the participants. For loud speech, participants were asked to keep a constant level of 80dB. The sustained vowel /a/ phonation was repeated three times per condition.

2.4. Data Processing and Analysis

Since the RIP and EMA recordings started asynchronous, we aligned the audio tracks of the EMA and RIP by an acoustic impulse at the beginning of the recording. The acoustic boundaries of both habitual and loud /a/ were manually segmented using Praat (Boersma & Weenink, 2024). For the EMA system, different distances between sensors were calculated and analyzed in the vertical (low-high, y) and horizontal (front-back, x) dimension (**Figure 2**):

- D1: Distance of the chest's middle sensor to the reference sensor (chest mid to R) → EMA_{D1}
- D2: Distance of the calculated midpoint between left sensor and right sensor on the chest to the reference sensor (midpoint to R) → EMA_{D2}
- D3: Distance of sternum to the reference sensor (sternum to R) → EMA_{D3}

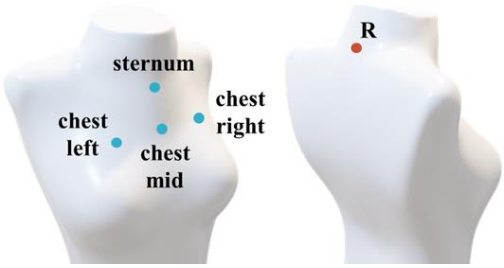


Figure 2: *Schematized EMA sensors on the front and on the back (R = reference sensor).*

For the calculated distances, three landmarks were automatically determined in the RIP and the EMA signal: (i) inhalation onset, (ii) inhalation peak, and (iii) exhalation offset (**Figure 3**). The landmark detection was as follows: The signals were prepared first by resampling them to 100 Hz and applying a Savitzky-Golay filter using a window of 101 samples and polynomial order 3 afterwards. The basis for the landmark detection was then the processed signal within a window of the acoustic boundaries of the target vowels ± 7s.

The signals' velocity was used for the detection of the inhalation onset and the exhalation offset. For the inhalation onset, the maximum velocity left to the inhalation peak was determined first and then the first zero crossing in the velocity was used for the landmark detection of the onset. The detection of the offset was based on the velocity multiplied by a window function consisting of two half Gaussians and a stable region during the acoustic segment. The last zero crossing left to the velocity maximum in the second half of the window was used as the offsets' landmark. The inhalation peak was defined as the maximum in the signal.

Figure 3 displays examples of synchronized RIP and EMA data during the production of sustained /a/ in habitual speech, namely the raw filtered signal, the resampled and filtered signal, the signals' velocity and the windowed velocity, along with the detected landmarks in vertical dashed lines.

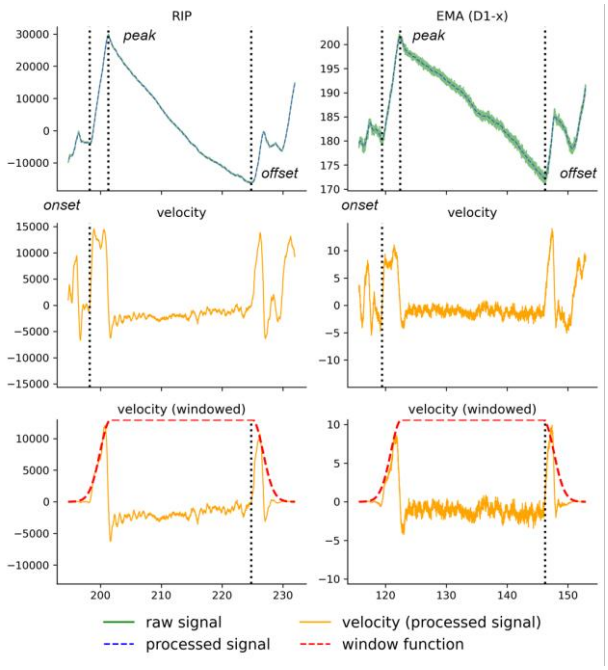


Figure 3: *Example of landmark detection in RIP (left) and EMA signal* $EMA_{D1}x$ *(right). Vertical dotted lines refer to landmarks (onset, peak and offset). Rows show the raw filtered and the processed signal (top), the velocity (mid), and the window function (bottom).*

To compare the RIP and EMA signal and to determine which EMA distance trajectories are most comparable to the RIP system, the procedure was as follows:

First, the following two parameters were calculated to analyze temporal breathing patterns:

1) Inhalation phase (s): Interval between inhalation onset and inhalation peak.
2) Exhalation phase (s): Interval between inhalation peak and inhalation offset.

To compare each of the two parameters, we run hierarchical Bayesian regression models for the two temporal parameters and speaking styles (loud, habitual) with the SIGNAL TYPE (RIP vs. $EMA_{D1}x$, $EMA_{D1}y$, $EMA_{D2}x$, $EMA_{D2}y$, $EMA_{D3}x$, $EMA_{D3}y$) as independent variables with by-speaker intercepts and slopes. We used default priors in all models. Results are reported under section 3.1.

Second, we compared the RIP and EMA trajectories based on 100 equally distanced time points from the inhalation onset to the exhalation offset and standardized the trajectories by token and signal type. For visual inspection, we calculated Euclidean-distance matrices showing the (dis-)similarity between RIP and the EMA dimensions across speakers and repetitions, and speaking styles (section 3.2.).

Third, we run Gaussian Process regression models for each speaking style on a subset of the standardized signal trajectories (steps of 5% from inhalation onset to exhalation offset). We used separate covariances for each SIGNAL TYPE with exponential priors for amplitude (lambda=1) and length scale (lambda=3) and a by-SIGNAL TYPE intercept with a default prior. The models were run with 2000 samples for tuning and

2000 samples in four chains, thus leading to 8000 iterations for the analysis. We computed the difference between the posterior of the RIP and the posterior of each EMA distance afterwards. Results are reported under section 3.3. We report the mean and the 95% highest density interval (HDI) of the posterior estimates for all regression analyses.

3. Results

3.1. Parameter comparisons

Table 1 contains the averaged results for the parameters of interest for the different signals (RIP vs. $EMA_{D1\text{-}D3}$) in both the x- and the y-dimension.

Table 1: *Mean durations of inhalation and exhalation phases in seconds (standard deviations in brackets) for the RIP and EMA distance signals.*

Condition	Signal	Inhalation phase	Exhalation phase
habitual	RIP	2.78 (1.18)	22.62 (8.50)
	$EMA_{D1}x$	2.36 (1.00)	22.49 (8.68)
	$EMA_{D1}y$	2.58 (1.02)	22.63 (8.63)
	$EMA_{D2}x$	2.81 (1.05)	22.49 (8.69)
	$EMA_{D2}y$	2.53 (1.07)	22.68 (8.59)
	$EMA_{D3}x$	2.29 (1.22)	22.10 (8.66)
	$EMA_{D3}y$	2.58 (1.02)	22.53 (8.59)
loud	RIP	2.41 (0.93)	23.46 (10.15)
	$EMA_{D1}x$	2.15 (0.99)	22.61 (10.37)
	$EMA_{D1}y$	2.14 (1.00)	23.23 (10.42)
	$EMA_{D2}x$	2.18 (1.02)	23.07 (10.59)
	$EMA_{D2}y$	2.15 (0.98)	23.36 (10.59)
	$EMA_{D3}x$	2.25 (1.10)	23.21 (10.17)
	$EMA_{D3}y$	2.22 (0.97)	23.16 (10.68)

No durational differences in the exhalation phases of the EMA signal (and its related differences) compared to RIP's in the production of sustained vowel /a/ in habitual and loud speech were found. However, regarding the inhalation phases, the models reveal slightly shorter inhalation phases in $EMA_{D1}x$ (β=-0.96 [-1.6, -0.35]) and $EMA_{D2}x$ (β=-0.45 [-0.82, -0.09]) in habitual and $EMA_{D1}x$ (β=-0.5 [-0.79, -0.18]), $EMA_{D2}x$ (β=-0.32 [-0.59, -0.05]) and $EMA_{D3}y$ (β=-0.35 [-0.67, -0.3]) in loud speech compared to the RIP signal.

3.2. Distance plots for visual inspection

Figure 4 and **Figure 5** display distance plots comparing RIP and EMA signals averaged across all speakers during sustained vowel productions in habitual speech (**Figure 4**) and loud speech (**Figure 5**). For the signal comparison in habitual and loud speech, the EMA_{D2y} signal was chosen as an example, as this EMA distance signal is most similar to the phases of the RIP signal - particularly in habitual speech (**Table 1**). The color coding indicates the continuum from similar (black; 0 of the normalized Euclidean distance) to dissimilar (white, 1 of the normalized Euclidean distance). 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 indicating a clear similarity between the trajectories of RIP and EMA.

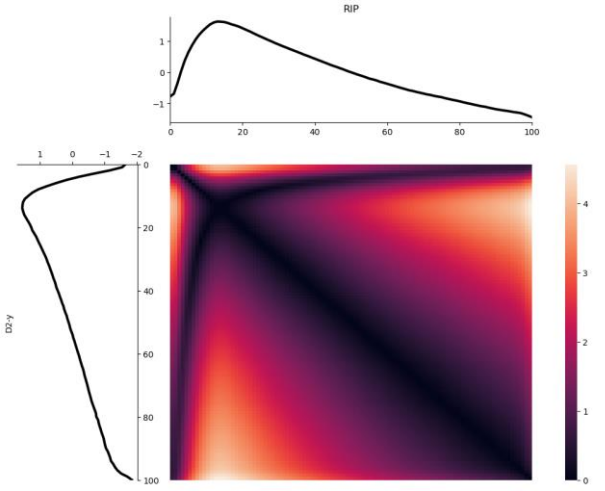


Figure 4: *Distance plot (EMA_{D2y}) comparing RIP and EMA signals in habitual speech.*

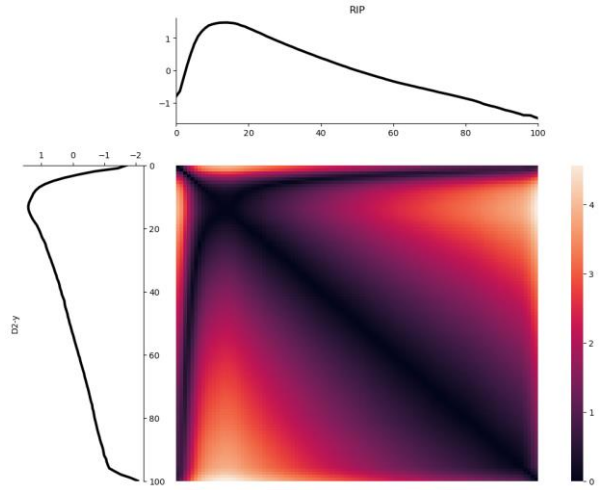


Figure 5: *Distance plot (EMA_{D2y}) comparing RIP and EMA signals in loud speech.*

3.3. Trajectory comparisons: Regression analysis

To investigate which distance signal is most suitable to track speech breathing patterns with EMA, we compare the contours of the RIP signal with all EMA distance signals by means of Gaussian Process regression models. **Figure 6** shows the output of the models for habitual (left column) and loud (right column) speech. Each panel shows the comparison of the RIP signal with the respective EMA signal. The top of each panel depicts the 95% posterior estimate for the RIP (blue, hatched) and the EMA signal (red), and the plot below shows the difference (orange) between the RIP signal and the EMA signal.

Our regression analyses revealed that none of the EMA distance signals significantly differs from the RIP signal in shape across the speech breathing movements. As can be seen in **Figure 6**, the 95% HDI of the posterior differences between the RIP and EMA contours is centered around zero, thus indicating no difference at each of the evaluated time points. If a significant deviation between the signals was detected, this would be marked by a red area (which is not the case here).

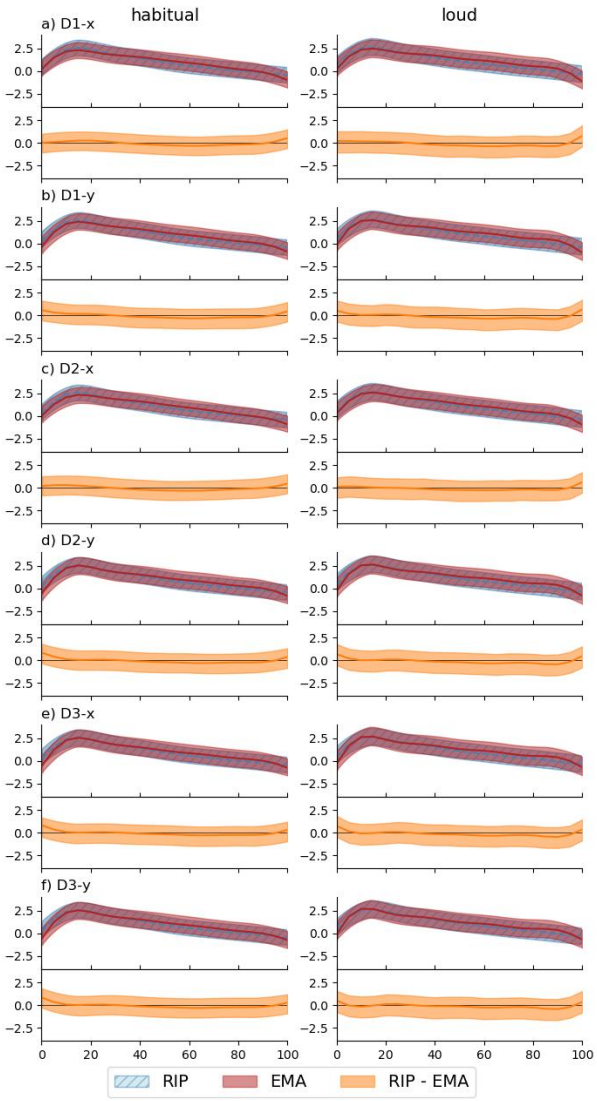


Figure 6: *Regression results for RIP compared to various EMA signals (rows) for habitual (left) and loud speech (right). The top of each panel shows 95% of the posterior estimates for RIP and the EMA signal, and the lower plot shows the difference between RIP and the EMA signal.*

4. Discussion

This study reveals that EMA sensors are capable of tracking speech breathing patterns that are comparable with the commonly used RIP signal. We were able to show that temporal parameters, such as inhalation and exhalation phases do not differ between the EMA and RIP signal. However, slightly longer durations were detected for some parameters. This could be explained by the fact that the expansion of the RIP band is measured in a three-dimensional space, whereas the EMA signal only measured one-dimensional distances. As EMA also allows for the analysis of 3D movement patterns, possible parameters need to be developed to capture 3D patterns in the future. Nonetheless, since the movement trajectories did not differ between RIP and EMA, we postulate that EMA is a potential method to collect speech breathing data.

As we attached EMA sensors to various positions on the chest, we were able to show that in principle, the signal from all sensors can be used. A subsequent analysis will determine which sensors are most suitable to give a recommendation on the minimum number of EMA sensors that should be used in future studies. In general, 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).

Due to the significant cost difference between an EMA system and an RIP, laboratories that already possess an EMA device can derive practical advantages from utilizing EMA instead of the traditionally employed RIP. The experimental process becomes simplified since there is no longer a requirement for diverse belts (as for Inductotrace®), resulting in enhanced convenience and reduced intrusiveness.

We will pursue the analyses of speaker-specific behaviors and look more into natural speech production, such as sentence productions and text reading.

5. Conclusion

Previous research has demonstrated that the respiratory inductive plethysmography (RIP) is a widely accepted and validated tool for studying speech breathing patterns. However, it also has its limitations, such as potential discomfort for participants and the possibility of body movements generating artifacts in the signal. This study is the first comparing speech breathing patterns assessed with Electromagnetic Articulography (EMA) to RIP signals. Results underscore the benefits and ease of using EMA for analyzing speech breathing pattern and paves the way for further studies which are using EMA systems to also easily collect data on speech breathing simultaneously to speech production kinematics.

6. Acknowledgements

This work has been carried out within the framework of the ANR-23-CE28-0017 INSPECT supported by the French ANR. Further, this work is partially supported by a public ANR grant as part of the program “Investissements d’Avenir” (ANR-10-LABX-0083). It contributes to the IdEx Université de Paris - ANR-18-IDEX-0001.

7. 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 timing in Hindi CV sequences

Shihao Du[1], Indranil Dutta[2], Adamantios I. Gafos[1]

[1]*Universtität Potsdam, Potsdam, Germany*
[2]*Jadavpur University, Kolkata, India*

shihao.du@uni-potsdam.de, indranildutta.lnl@jadavpuruniversity.in, gafos@uni-potsdam.de

Abstract

This short report presents some preliminary results from electromagnetic articulography (EMA) recordings of Hindi Consonant-Vowel (CV) sequences. We specifically asked if and how articulatory timing in CV, quantified by the interval from V-target to C-offset, is modulated by consonant phonation, consonant place of articulation and vowel quality. Results show that vowel height and frontness and C place of articulation exert significant effects on CV timing, whereas C phonation (voicing and aspiration) has no significant effect on the interval we chose to quantify CV timing here. Potential explanations of the disparity between vowel-related and consonant-related effects are suggested.

Keywords*: speech production, Consonant-Vowel articulatory timing, Hindi*

1. Introduction

In stop-vowel sequences of Hindi, we compared Consonant-Vowel (CV) timing for different CVs where the consonant was one of /b/, /p/, /bʱ/, /d/, /t/, /dʱ/, /tʰ/ and the vowel was one of /iː, ɪ, uː, ʊ, eː, e, oː, o, aː/. Not much is known about how phonation and place of articulation of the consonant affects articulatory timing in CV sequences. Likewise, much is yet to be documented on how properties of the vowel affect its timing with its preceding consonant. Early studies such as Ostry et al. (1983) and Löfqvist and Gracco (1997) on English did report on a possible consonant voicing effect (but whether the effect was due to voicing or aspiration could not be determined given the language) and effects related to vowel quality on the kinematics of the consonantal gestures, but how these features jointly affect CV timing remains largely unknown. For consonant sequences, Bombien and Hoole (2013) show that the temporal distance between the oral constrictions in German stop-liquid sequences (e.g., [gl] versus [kl]) varies systematically as a function of stop voicing. The former shows about 21 ± 2 ms more overlap than the latter. Whereas these studies focus on effects of voicing on the timing of the C oral gestures in CC sequences, our study focuses on the timing between the oral gestures of C and V in CV sequences. Our motivation is the same as that in Bombien and Hoole (2013) who note that "the coordination of supra-laryngeal articulations with respect to laryngeal specification is an area of speech production research which so far has received only limited attention and is far from being understood" (Bombien & Hoole 2013: p. 539). Hindi offers an ideal case study in this respect. In the CV context, where C is a stop, consonants exhibit a four-way contrast (in alveolar, retroflex, and velar stops; labial stops show primarily a three-way contrast, as /pʰ/ is realized increasingly as [f]), with the full suite of voiced unaspirated, voiced aspirated (also known as breathy), voiceless unaspirated, and voiceless aspirated stops.

We give an example of how the lack of knowledge in this domain has hindered theory development and evaluation. Browman and Goldstein (1988) first observed that when adding a consonant to the start of a syllable, from [pa] to [spa], the temporal organization of the whole changes such that [p], [a] timing in [spa] is different from that in [pa]. The gestures of [p], [a] seem to slide closer to one another in [spa] than in [pa]. It was hypothesized that the vowel onset in such sequences is synchronous with the center of the prevocalic consonantism (be it a single [p] or an [sp]) and specifically with the midpoint of the consonantal closure intervals of all consonants (Browman & Goldstein 1988: p. 150; see also Honorof & Browman 1995, Figure 1, p. 552). As the American English stop in an [s]-stop cluster before a vowel is not aspirated (but the lone voiceless stop is), such a comparison implies a potential confound (see also Katz 2012) due to the phonation (presence versus absence of the aspiration gesture) which may independently affect vowel timing. Perhaps a more appropriate comparison would be to consider the timing of the vowel in relation to the prevocalic consonantism in [s]-stop-vowel versus single voiced stop-vowel sequences, because in both the stop is not aspirated; this is still imperfect, however, because of the presence of the /s/ which makes it impossible to decide whether any differences are exclusively due to the phonation of the stop (because /s/ also implicates tongue movement just like the vowel following the stop, it may be that whatever requirements /s/ imposes on tongue body control, these have an influence on the timing of the subsequent vowel which also implicates the tongue body). In any case, the facts are simply not known here. An ever more appropriate comparison would be to compare the timing of the vowel in relation to the prevocalic C in single, not aspirated stop-vowel sequences versus single aspirated vowel sequences but the former are not available in English.

Consider furthermore the fact that typically segments are ensembles of gestures. In defining the notion of inter-segmental coordination, which gestures from the segments so coordinated are to be related to one another? Is the glottal opening gesture of a [t] or the velic lowering gesture of an [m] eligible for entering in a coordination relation with other segments? In Gafos (2002), inter-segmental coordination was defined by making reference to notion of 'head' of a segment: "Two segments S1, S2 are coordinated with some coordination relation λ, /S1 λ S2/, if the head gestures of these segments are coordinated as in λ" (Gafos 2002: 284), where coordination was operationalized by specifying that one landmark from the first and another from the second gesture are aligned in time (synchronized). The head gesture of a segment is the gesture of the oral task variable of that segment (Browman & Goldstein 1986; Saltzman 1986). This can be motivated on a number of reasons. Theoretical precedent in feature-geometric representations pointed to the key role of the oral gesture of a segment (Sagey 1986; Halle 1995). Kingston's (1985) work on "articulatory binding" proceeded from the fact that contrastive laryngeal articulations tend to be bound to the release of oral stops. Steriade (1993; 1994) formulated a theory of representations which directly encoded so-called "anchor" positions of oral closure and release to explain facts about possible segments with contrastive laryngeal and velic specifications. It was on the backdrop of these proposals that oral gestures were assumed to drive segment-to-segment coordination. Finally, the data Gafos (2002) aimed to account

for indicated that laryngeal or velic gestures did not enter into the phonological and morphological effects that provided the core argument for a grammar of gestural coordination in that study. Thus, identity avoidance effects were observed for adjacent segments with identical oral gestures (e.g., [d-t]) but not so for identical velic or laryngeal gestures. An [n-m] or a [t]-[k] sequence did not trigger identity avoidance effects even though these are sequences of two identical velic lowering and laryngeal gestures respectively. It was on the basis of such facts that inter-segmental coordination relations were proposed to be stated by reference to the oral gestures of the segments so coordinated, with the intra-segmental laryngeal or velic gestures following suit by maintaining their segment-internal relation to the head gesture of their segment (i.e., when the oral gestures slide apart, their corresponding velic gestures slide along with them). If inter-segmental coordination in CV sequences is not mediated by laryngeal specifications of the consonant, this implies that CV timing in Hindi should not be modulated by the phonation characteristics of the C (voiceless unaspirated, voiceless aspirated, voiced unaspirated, and voiced unaspirated stops). It is thus clear that further theory evaluation and development rely crucially on a better understanding of the facts regarding the role of consonant phonation and place of articulation on the timing of the oral gestures in CV sequences.

Recently, intervals delineated by landmarks on CV sequences have been examined in works that aim to assess the extent to which inter-segmental coordination can be expressed in terms of synchronicity relations among landmarks. For instance, Shaw and Chen (2019) demonstrated on basis of Mandarin CV sequences consisted of labial consonants (/m/ and /p/) and back rounded vowels (/ou/, /u/, /uo/) that the lag from V-target to C-offset has a mean of zero, representing close synchrony of the two landmarks. In another study along the same lines, 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). Similarly, Durvasula and Wang (2023) examined whether it is V-onset or V-target that is aligned to some landmark within the prevocalic consonantal gesture in five American English words (*back*, *fiber*, *make*, *much*, *people*) with a word-initial labial obstruent-vowel sequence and reported that V-target was consistently aligned with the C-offset. 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.

2. Methods

Electromagnetic articulography data were collected from 2 native male speakers of Hindi aged from 22 to 23, who reported no hearing or other health issues. The speakers produced 63 target words beginning with CV sequences where the consonant was either /b/, /p/, /bʱ/, /d/, /t/, /dʱ/, or /tʰ/ (aspirated /pʰ/ is not included because in Hindi it underwent fricativization and is realized contemporarily as [f]) and the vowel was one of /ˈiː, ɪ, uː, ʊ, eː, e, oː, o, aː/ (the consonants and vowels were fully crossed, such that each consonant is paired with all nine vowels; i.e., 7 consonants × 9 vowels = 63 CV sequences). The target words were all embedded in the carrier phrase *Ramā* ___ *bolī* (translation: 'Ramā said ___'), in which the target appears at a phrase-medial and prosodically neutral position. Each phrase was repeated 10 times by each speaker in a random order, yielding 1260 tokens in total (63 target words × 10 repetitions × 2 speakers). The Carstens AG501 device was used to record movements of 10 sensors attached to the speech organs and head at a sampling rate of 1250 Hz. For the current study, the movements of the sensor attached to the tongue dorsum (TD) was used to identify vowel gestures, the sensor attached to the tongue tip (TT) for the gestures associated with the alveolar consonants, and the Euclidean distance between the sensors attached at the vermillion border of the upper and lower lips (UL and LL) for the bilabial consonant gestures. Articulatory gestures were parsed manually using the matlab-based software MVIEW (Tiede 2005). Temporal landmarks were identified using a 20% peak velocity threshold. Out of the elicited 1260 tokens, 77 tokens (6.11%) were eliminated because of data storage failure or failure of gestural parsing. For each of the remaining tokens, the temporal distance from consonant offset and vowel target was computed to assess landmark synchrony in CV sequences. A linear-mixed effects model was fitted to the data with the synchrony measure as the dependent variable and consonant voicing (voiced vs. voiceless), aspiration (aspirated vs. un-aspirated), place (alveolars 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.

3. Results

We first set out to assess the extent to which pairs of landmarks drawn from the vowel and the consonant, such as the landmarks V-target and C-offset, show synchrony. **Table 1** below lists means and standard deviations for the intervals C-onset to V-onset, V-onset to C-target, C-target to V-target, V-target to C-offset as well as the inter-plateau interval (C-release to V-target) and the interval from C-opening peak velocity (PV) to V-target. It can be seen that the V-target to C-offset interval has a mean of 8.54 ms in our Hindi dataset, which is the mean closest to zero among all the tested intervals, indicating near synchrony. This result is in line with findings from other recent work (Shaw & Chen 2019; Kramer et al. 2023; Durvasula & Wang 2023) which suggests that V-target and C-offset may be (near) synchronous in CV timing.

Table 1: *Means and standard deviations of six intervals delineated by a landmark on the consonant and a landmark on the vowel in Hindi CV sequences.*

Interval	Mean (ms)	SD (ms)
C-onset to V-onset	136.13	48.69
V-onset to C-target	62.81	45.43
C-target to V-target	156.07	41.50
C-release to V-target	106.10	36.60
C-opening-PV to V-target	53.55	37.87
V-target to C-offset	8.54	44.66

We then assessed how consonant phonation, place of articulation, and vowel quality modulate the duration of this interval. **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 alveolar 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. Finally, with regard to frontness / backness, back rounded vowels had 10.9 ms longer lag than non-back unrounded vowels.

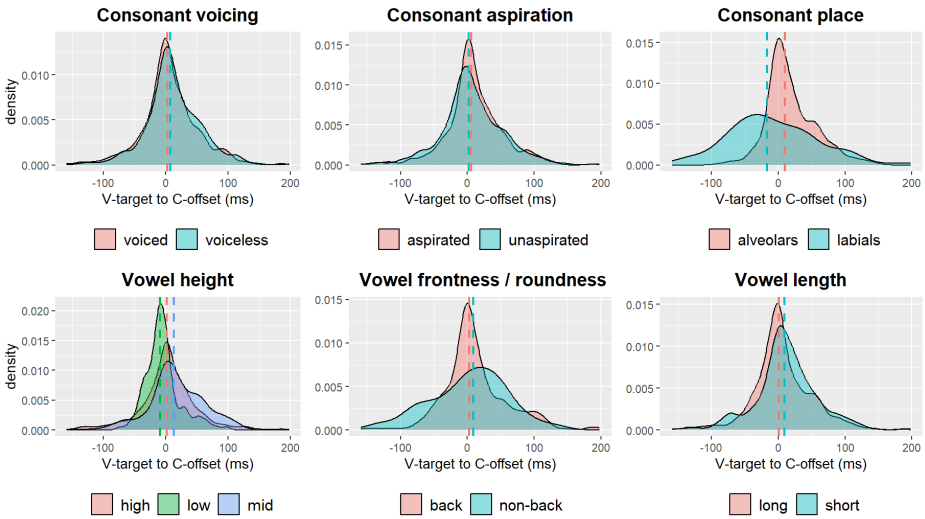


Figure 1: *Distribution of the V-target to C-offset interval across subjects as a function of consonant voicing, aspiration, place, vowel height, frontness / roundness, and length. Vertical lines are the medians in each group.*

Table 2: *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.*

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***

4. Discussion and conclusion

A main result emerging from our data is that 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). Why may this be so? Early studies on English CV sequences (Ostry et al. 1983, Löfqvist and Gracco 1997) reported robust effects of vowel quality on the consonant's kinematics, with any effects of consonant voicing being place-specific or not consistent across subjects. Thus, Löfqvist and Gracco (1997) reported no consistent voicing effect in labial consonant-initial CVs (their stimuli consist of only labials), 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 (their stimuli consist of only velars). 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 phonation 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 2** 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 limited compared to vocalic effects in terms of their number (3 significant aspiration and voicing effects vs. 8 significant height

and frontness effects), but 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 closing 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.

In conclusion, it has been found that vowel height and frontness and C place of articulation exert significant effects on 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 the transition between the C and the V. This may then explain the presence of vowel quality effects and the absence of consonant phonation effects in CV timing. Of course, our preliminary results are limited, given the choice to quantify CV timing in the specific way chosen here, which is motivated by recent work reporting on this interval (Shaw & Chen 2019; Kramer et al. 2023; Durvasula & Wang 2023).

We note that despite the fact that vowel quality significantly affects extent of landmark synchrony, the V-target to C-offset interval also shows a relatively high standard deviation as documented in **Table 1** (though not the highest as in the results reported by Kramer et al. 2023), implying that it is not the most stable interval. The more extensive set of CV sequences examined in our data compared to earlier work brings out the specificity of such descriptive statistics (on interval variability) as a function of segmental composition. In turn, these results indicate that taking grand means of these intervals across all CV sequences may not be appropriate given the significant effects of V quality in our data. Moreover, the fact that V-target to C-offset interval is relatively variable both in our data as well as in the data from Kramer et al. (2023) hints at the insufficiency of considering synchrony alone as the sole basis of inter-gestural coordination (as noted in Kramer et al. 2023).

5. Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 317633480 – SFB 1287.

6. References

Bombien, L., & Hoole, P. (2013). Articulatory overlap as a function of voicing in French and German consonant clusters. *The Journal of the Acoustical Society of America*, *134*(1), 539–550. https://doi.org/10.1121/1.4807510

Browman, C. P., & Browman, L. M. (1986). Towards an Articulatory Phonology. *Phonology Yearbook*, *3*, 219–252. JSTOR.

Browman, C. P., & Goldstein, L. M. (1988). Some notes on syllable structure in Articulatory Phonology. *Phonetica*, *45*(2–4), 140–155. https://doi.org/10.1159/000261823

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.

Halle, M. (1995). Feature geometry and feature spreading'. *Linguistic Inquiry* 26, 1–46.

Honorof, D. N. and Browman, C. P. (1995). The center or edge: how are consonant clusters organized with respect to the vowel. In Elenius, K. and Branderud, P. (Eds). Proceedings of the 13th ICPhS, Stockholm, Sweden, volume 3, pages 552-555.

Katz, J. (2012). Compression effects in English. *Journal of Phonetics*, *40*(3), 390–402. https://doi.org/10.1016/j.wocn.2012.02.004

Kingston, J. (1985). The phonetics and phonology of the timing of oral and glottal events. [Ph.D. dissertation]. University of California, Berkeley.

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

Nam, H., & Saltzman, E. (2003). A competitive, coupled oscillator model of syllable structure. Proceedings of the 15th International Congress of Phonetic Sciences, 2003, 2253–2256.

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

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: a preliminary model. In H. Heuer & C. Fromm (Eds.), *Generation and Modulation of Action Patterns* (pp. 129–144). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-71476-4_10

Sagey, E. (1986). The representation of features and relations in non-linear phonology. [Ph.D. dissertation]. MIT. [Published 1991, Garland, New York.]

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

Steriade, D. (1993). Closure, release, and nasal contours', in Huffman, M. K. and Krakow, R. A. (Eds.), *Phonetics and Phonology 5: Nasals, Nasalization, and the Velum, Academic Press*, New York, pp. 401–470.

Steriade, D. (1994). Complex onsets as single segments: the Mazateco pattern', in Cole, J. and Kisseberth, C. (Eds.), *Perspectives in Phonology*, CSLI, Stanford, pp. 203–291.

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

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]*Unidade Local de Saúde de Coimbra, 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

Abstract

The purpose of this study was to define acoustic cues used to discriminate vowel height in voiced and whispered speech. Seventeen speakers produced sustained oral vowels, disyllabic words, sentences and read a phonetically balanced text. These tasks were repeated in voiced and whispered speech and analysed using the following parameters: Fundamental frequency, formant frequencies, spectral slope, sound pressure level and durations. Kernell density estimation plots and Pillai scores were used to characterise the vowel spaces and the degree of overlap between vowels. First formant frequency and relative duration were consistently used as height cues across the two speech modes (voiced and whispered). Whispered vowel spaces shifted downward (relative to voiced), and vowel pairs /i-a/, /a-ɔ/, /ɔ-u/ and /u-i/ were distinct when produced in both speech modes. The evidence presented can be used to restore voiced speech signals and to inform rehabilitation strategies.

Keywords*: speech production, whispered speech, vowels*

1. 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).

Some of this work has been previously published as part of an open access paper (Jesus et al., 2023).

2. 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 Sennheiser Ear Set 1 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.

Kernell density estimation plots were used to characterise the vowel spaces. They resemble "topographic maps of hills" with density information that "allows the quick identification of central tendencies and possible bimodal distributions needing further inspection". They "work for sparse, skewed, or imbalanced data" (Freeman, 2023), such as ours.

The degree of overlap between /i-a/, /a-ɔ/, /ɔ-u/ and /u-i/ vowel pairs was quantified with Pillai scores, because they have been recently shown (Freeman, 2023; Stanley & Sneller, 2023) to model better vowel categories "than other methods due to their ability to account for multiple dimensions, skewed distributions, unequal densities, and sparse data" (Freeman, 2023). The null hypothesis of the Multivariate Analysis of Variance from which the Pillai scores were generated (Stanley & Sneller, 2023, p. 57) was that the two vowels overlap.

Two mixed effects regression models were developed using the `lmer` function from the lme4 version 1.1-33 package, both models with the `pillai_score` as outcome variable (Li et al., 2023, p. 1179), one model considering `vowel_pair` as a fixed effect, `speaker_id` and `speech_mode` as random effects, and an additional model considering `speech_mode` as a fixed effect, `speaker_id` and `vowel_pair` as random effects. Results from likelihood ratio tests of the models with the `vowel_pair` and `speech_mode` effects against the models without the `vowel_pair` and `speech_mode` effects were also analysed.

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. The models' predictions and lines spanning the 95% confidence interval were drawn using the sjPlot 2.8.14 package.

3. 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.

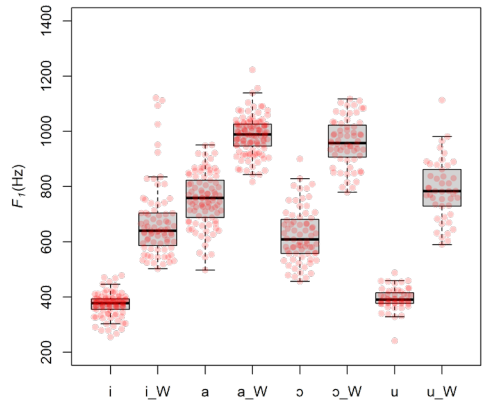


Figure 1: *Female voiced /i, a, ɔ, u/ and whispered /i_W, a_W, ɔ_W, u_W/ vowels' F_1 frequencies in a phonetically balanced text.*

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 $F_{1(\text{whispered})}$ - $F_{1(\text{voiced})}$ of female speakers (Pearson's correlation coefficient = 0.660, p = 0.005; two-tailed p-value).

3.1. Vowel spaces

The vowel space areas analysed using kernel density plots, revealed a compression in whispered speech, when compared to an equivalent voiced speech task, both for female (shown in **Figure 2**) and male (shown in **Figure 3**) speakers. A clear downward shift (relative to voiced speech) of vowel spaces in whispered speech could be observed for all speech tasks.

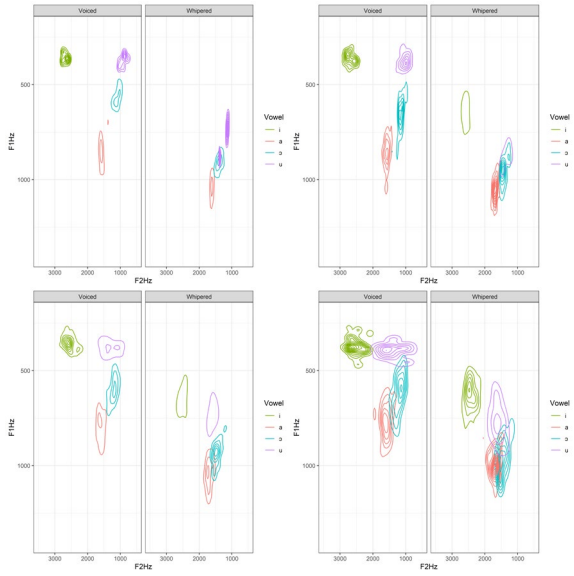


Figure 2: *Kernell density plots of female's vowels formant frequencies (sustained – top left; words – top right; sentences – bottom left; text – bottom right).*

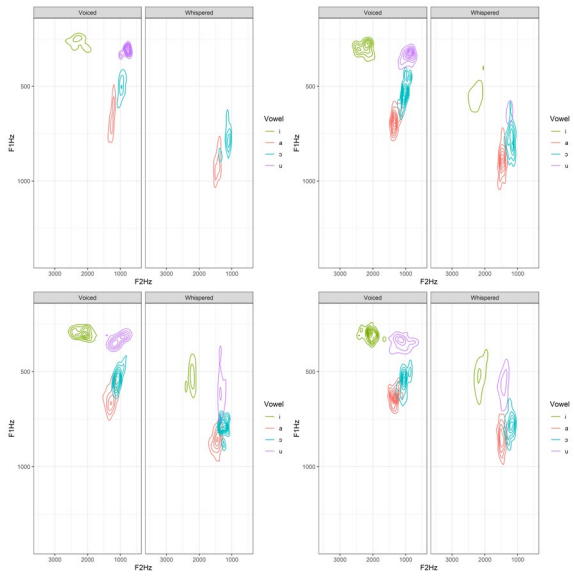


Figure 3: *Kernell density plots of male's vowels formant frequencies (sustained – top left; words – top right; sentences – bottom left; text – bottom right).*

Kernel density plots of the Pillai scores for all voiced and whispered vowels, shown for females in **Figure 4**, revealed that as the speakers produced vowels in a more natural task (sustained → words → sentences → text) the estimates of the

probability density functions of voiced and whispered speech were more alike.

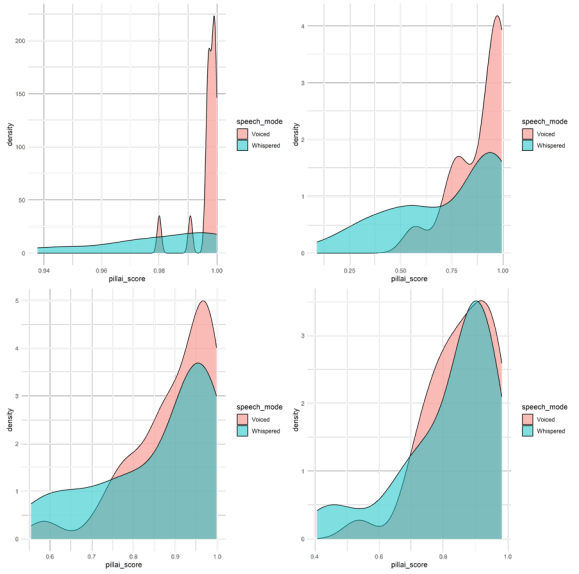


Figure 4: *Kernel density plots of the Pillai score differences between voiced and whispered vowels (sustained – top left; words – top right; sentences – bottom left; text – bottom right).*

Likelihood ratio tests of the mixed effects regression model `pillai_score ~ speech_mode + (1|speaker_id) + (1|vowel_pair)` with the `speech_mode` effect against the model without the `speech_mode` effect, only revealed a significant difference between models for female's sustained vowels, and for both female's and male's vowels in words. That is, there was a significant difference between Pillai scores of the voiced and whispered vowels (lower for whispered speech, i.e., vowels were more overlapped for whispered speech) in a very limited (less natural) number of tasks: Female (sustained) – $\chi^2(1) = 11.97$, $p = 0.001$; Male (sustained) – $\chi^2(1) = 0.62$, $p = 0.432$; female (words) – $\chi^2(1) = 14.83$, $p < 0.001$; male (words) – $\chi^2(1) = 9.53$, $p = 0.002$; female (sentences) – $\chi^2(1) = 3.68$, $p = 0.055$; male (sentences) – $\chi^2(1) = 0.18$, $p = 0.669$; female (text) – $\chi^2(1) = 3.41$, $p = 0.065$; male (text) – $\chi^2(1) = 0.47$, $p = 0.492$.

An individual analysis of Pillai scores for the vowel pairs /i-a/, /a-ɔ/, /ɔ-u/ and /u-i/ produced by female and male speakers in voiced and whispered words, revealed higher values than a threshold, calculated using a formula recently proposed by Stanley and Sneller (2023, p. 61) as a standard for quantifying mergers in sociolinguistics (Grama, 2023), and the *p*-values were less than 0.05. There were, however, some exceptions, i.e., there was an overlap between /a-ɔ/ in sentences and text produced by two male speakers, and the productions of whispered /ɔ-u/ and /u-i/ in sentences and text by nearly all male speakers. For some speakers, it was not possible to calculate the Pillai scores due to a limited number of viable vowel exemplars for reliable formant estimation.

A mixed effects regression model with the lme4 syntax `pillai_score ~ vowel_pair + (1|speaker_id) + (1|speech_mode)` predicted the values shown in **figures 5** and **6**. Likelihood ratio tests of the model with the `vowel_pair` effect against the model without the `vowel_pair` effect revealed a significant difference between models, i.e., there was a significant difference between Pillai scores of the four vowel pairs: Female (sustained) – $\chi^2(3) = 12.43$, $p = 0.006$; male (sustained) – $\chi^2(3) = 18.15$, $p < 0.001$; female (words) – $\chi^2(3) = 47.17$, $p < 0.001$; male (words) – $\chi^2(3) = 59.67$, $p < 0.001$; female (sentences) – $\chi^2(3) = 50.98$, $p < 0.001$; male (sentences) – $\chi^2(3) = 44.22$, $p < 0.001$; female (text) – $\chi^2(3) = 45.60$, $p < 0.001$; male (text) – $\chi^2(3) = 58.62$, $p < 0.001$.

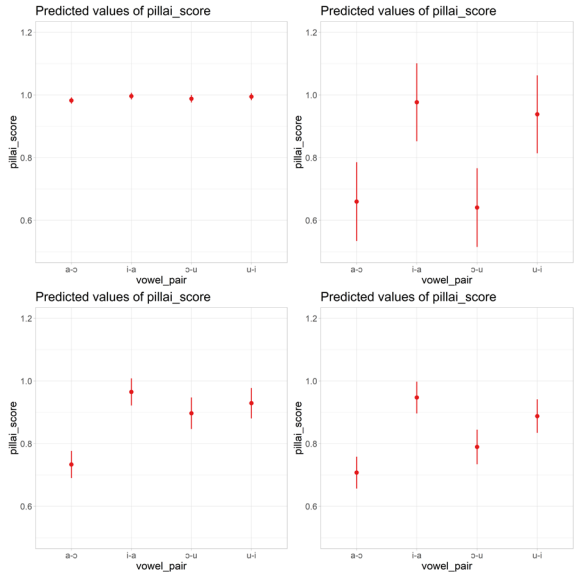


Figure 5: *Model predictions and lines spanning the 95% confidence interval for female vowels (sustained – top left; words – top right; sentences – bottom left; text – bottom right).*

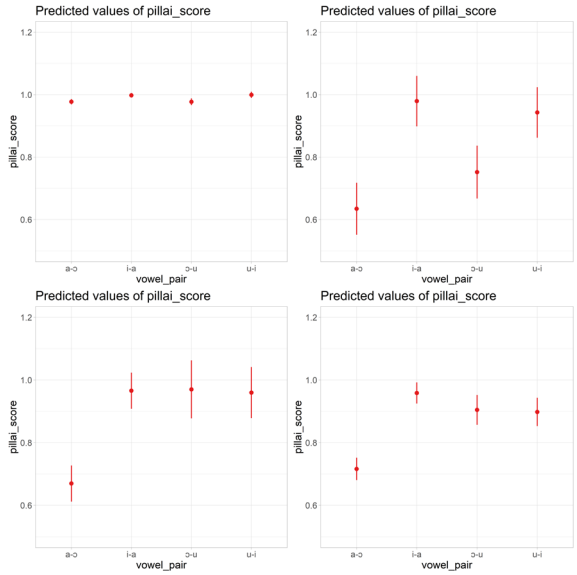


Figure 6: *Model predictions and lines spanning the 95% confidence interval for male vowels (sustained – top left; words – top right; sentences – bottom left; text – bottom right).*

3.2. Durations

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 7** correspond to the phone to text-length ratio of the text task.

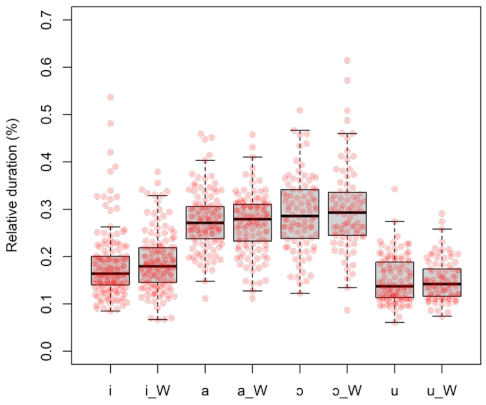


Figure 7: *Female voiced /i, a, ɔ, u/ and whispered /i_W, a_W, ɔ_W, u_W/ vowels' relative durations in a phonetically balanced text.*

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, even when the speakers whispered the vowels.

4. Discussion and conclusion

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_0 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.

We could also conclude that the vowel pairs were distinct, i.e., even the pairs /a-ɔ/ and /ɔ-u/ that were acoustically close in the vowel space were marginally contrastive in both speech modes (voiced and whispered). In the more natural tasks (sentences and text) the underlying distribution of the Pillai scores were not significantly different.

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).

This paper lays the groundwork for signal processing-based algorithms aiming at restoring voicing in whispered speech signals and can inform rehabilitation strategies.

5. Acknowledgements

This work was financially supported by Project PTDC/ EMDEMD/ 29308/ 2017 - POCI-01-0145-FEDER-029308 - funded by FEDER funds through COMPETE2020 - Programa Operacional Competitividade e Internacionalização (POCI) and by national funds (PIDDAC) through FCT/MCTES. Support was also received from National Funds through the FCT - Foundation for Science and Technology, in the context of the projects UIDB/00127/2020 and UIDB/04106/2020.

6. 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.

Freeman, V. (2023). Production and perception of prevelar merger: Two-dimensional comparisons using Pillai scores and confusion matrices. *Journal of Phonetics*, *97*. https://doi.org/10.1016/j.wocn.2023.101213

Grama, J. (2023). Change Over Time in [E] and [ae] in Hawaii Creole. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023), Prague, Czech Republic*, 2986–2990.

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

Li, P., Fledge, J. E., Martin, C. D., & Kartushina, N. (2023). Speech Stability Across Time: Evidence from Norwegian Vowels in Spontaneous Speech Production. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023), Prague, Czech Republic*, 1177–1181.

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.

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

Stanley, J. A., & Sneller, B. (2023). Sample size matters in calculating Pillai scores. *The Journal of the Acoustical Society of America*, *153*(1), 54–67. https://doi.org/10.1121/10.0016757

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

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]*Unidade Local de Saúde de Coimbra, 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

Abstract

Not much is known about acoustic cues concerning consonant place and voicing of whispered fricatives, so the productions of sustained sibilants, disyllabic words, sentences and reading of a phonetically balanced text, were compared in voiced and whispered speech modes. Spectral peak frequencies and levels, spectral slopes, sound pressure level and durations were calculated. A Functional Principal Component Analysis (FPCA) of Power Spectral Density (PSD) estimates of voiced, whispered, and whistled fricatives was developed. The broad peak frequency was used to discriminate /s/ and /ʃ/. Spectral slope and broad peak frequency were associated with place. FPCA scores revealed various sources of variation. PSD was significantly different in voiced, whispered, and whistled fricatives. The relative duration of same-place voiceless fricatives was higher than voiced fricatives both in voiced and whispered speech. This evidence can be used to restore voiced speech signals from aphonic patients.

Keywords*: speech production, whispered speech, fricatives*

1. 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.

The fricatives analysed in this paper are produced with the tongue "forming the jet-producing constriction" (Shadle, 2010, p. 62), resulting in an unstable boundary layer at the teeth. A front cavity feedback mechanism can reinforce this instability generating a narrow-bandwidth peak (Shadle, 1983, 2010), observed in whistling.

Whistled articulations of both voiced and whispered speech have been described as "the use of an extremely narrow channel in target /s/ and /z/", "producing a whistling sound instead of the normal friction" (Ball & Local, 1996, p. 58). It is, however, possible to combine "tones and broadband noise" a mechanism that "could be at work with whistly fricatives" when an "unstable jet formed by an orifice", produced by raising the tongue, strikes the teeth (Shadle, 2010, p. 62).

Alveolar fricatives have been previously (Shadle & Scully, 1995, p. 64) identified as whistled using auditory perception and spectral analysis. Whistling could reinforce fricatives' source strength (Benninger et al., 1988), so the expected lower source strength in whispered speech might no longer be observed.

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).

2. 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 (PSD) 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.

Functional Principal Component Analysis (FPCA) was also used to explore variation (Cronenberg et al., 2020) of the PSD for voiced/ whispered pairs of sibilant fricatives produced by the 8 female speakers, and not whistled/ whistled alveolar fricatives pairs. PSD estimates were processed using a script written by Gubian (2023) based on two R packages develop by Happ-Kurz (2020): funData 1.3-8 and MFPCA 1.3-10. Both PC1 and PC2 scores (s1 and s2 shape descriptors), were used to linearly model (`lm` function in R) the curves using the following reconstruction formulas:

$$\text{predCurve}_{\text{voiced}}(f) = \mu(f) + s1_{\text{voiced}} \cdot PC1(f) + s2_{\text{voiced}} \cdot PC2(f) \quad (1)$$

$$\text{predCurve}_{\text{whispered}}(f) = \mu(f) + s1_{\text{whispered}} \cdot PC1(f) + s2_{\text{whispered}} \cdot PC2(f) \quad (2)$$

$$\text{predCurve}_{\text{not\_whistled}}(f) = \mu(f) + s1_{\text{not\_whistled}} \cdot PC1(f) + s2_{\text{not\_whistled}} \cdot PC2(f) \quad (3)$$

$$\text{predCurve}_{\text{whistled}}(f) = \mu(f) + s1_{\text{whistled}} \cdot PC1(f) + s2_{\text{whistled}} \cdot PC2(f) \quad (4)$$

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 and data visualisation.

3. Results

The parametric analysis of the PSD estimates revealed 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 non-parametric 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), in both in voiced and whispered speech modes, as shown in **Figure 1**.

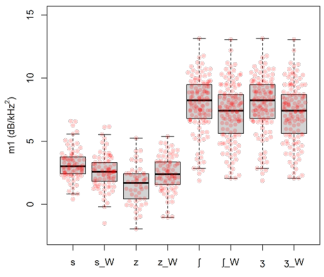


Figure 1: *Male voiced /s, z, ʃ, ʒ/ and whispered /s_W, z_W, ʃ_W, ʒ_W/ fricatives' low frequencies spectral slope (m1) in words.*

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; /ʒ/ 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 non-parametric comparison for post hoc testing after a Kruskal-Wallis tests; one-tailed p-values) than for postalveolar fricatives /ʃ, ʒ/, in both speech modes, in all four speech tasks and for both sexes, as shown in **Figure 2**.

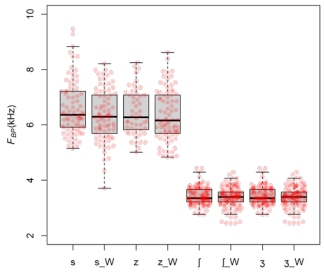


Figure 2: *Male voiced /s, z, ʃ, ʒ/ and whispered /s_W, z_W, ʃ_W, ʒ_W/ fricatives' broad peak frequency in words.*

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.

3.1. FPCA of Voiced and Whispered PSD

FPCA was used to explore the main dimensions of variation of the PSD estimates of voiced/ whispered pairs for all the fricatives produced by the 8 female speakers (shown in **Figure 3** for words).

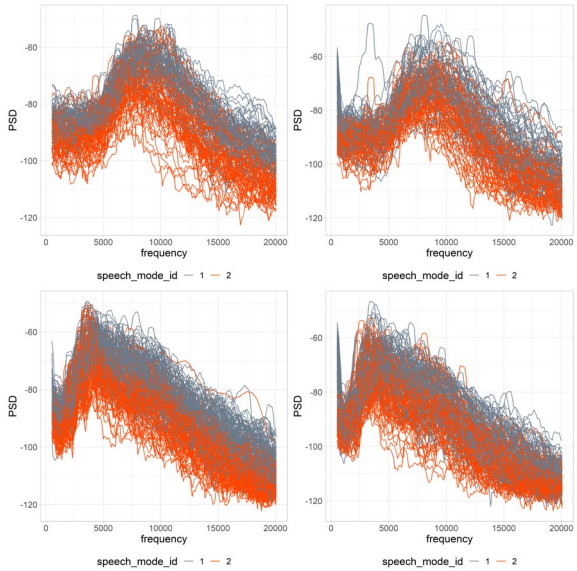


Figure 3: *PSD estimates in dB (frequency in Hz) for voiced (grey) and whispered (orange) fricatives. Top left /s, s_W/, top right /z, z_W/, bottom left /ʃ, ʃ_W/ and bottom right /ʒ, ʒ_W/.*

The FPCA curves for components PC1 and PC2, shown in **Figure 4** for words, had the following impact on the shape of the curves (proportion of explained variance): /s/ – PC1 = 93.3 %, PC2 = 6.7 %; /z/ – PC1 = 90.4 %, PC2 = 9.6 %; /ʃ/ – PC1 = 87.3 %, PC2 = 12.7 %; /ʒ/ – PC1 = 89.0 %, PC2 = 11.0 %.

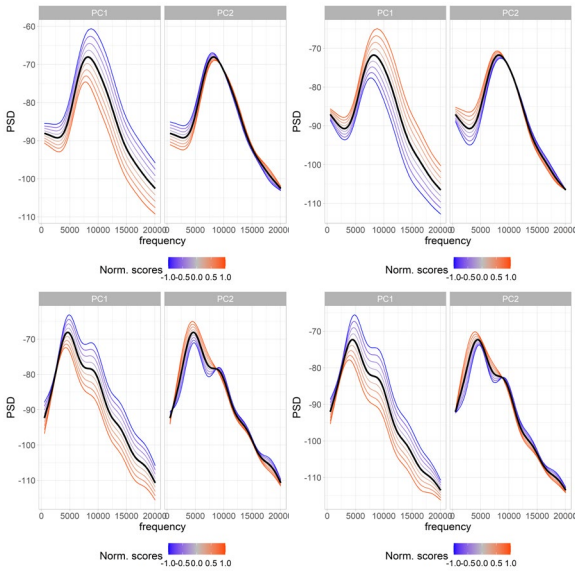


Figure 4: Voiced/ whispered *FPCA curves for components PC1 and PC2, and the effect of scores. Top left /s, s_W/, top right /z, z_W/, bottom left /ʃ, ʃ_W/ and bottom right /ʒ, ʒ_W/.*

The reconstructing of curves (shown in **Figure 5** for words) was based on s1 and s2 scores with the following standard deviations (sd): /s/ sd_{s1} = 969.55, sd_{s2} = 257.32; /z/ – sd_{s1} = 931.41, sd_{s2} = 299.57; /ʃ/ – sd_{s1} = 838.21, sd_{s2} = 318.63; /ʒ/ – sd_{s1} = 857.85, sd_{s2} = 301.06. The proportions of variance explained by the regression models predicting s1 were: /s/ – R^2 = 0.406, p = 0.000; /z/ – R^2 = 0.270, p = 0.000; /ʃ/ – R^2 = 0.486, p = 0.000; /ʒ/ – R^2 = 0.210, p = 0.000. The proportions of variance explained by the regression models predicting s2 were: /s/ – R^2 = 0.001, p = 0.668; /z/ – R^2 = 0.005, p = 0.422; /ʃ/ – R^2 = 0.012, p = 0.094; /ʒ/ – R^2 = 0.017, p = 0.073.

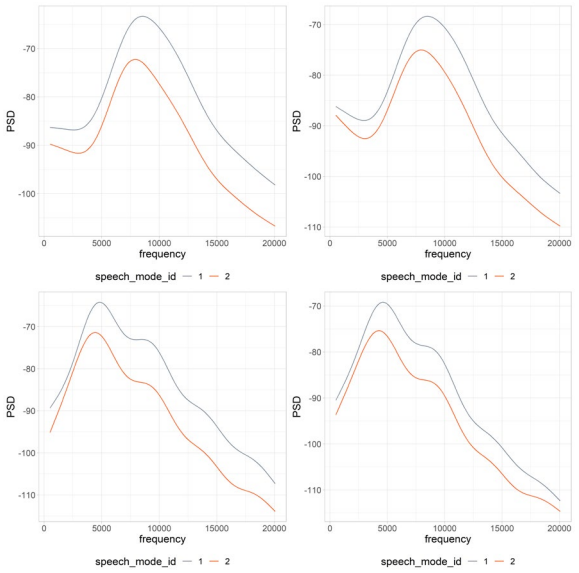


Figure 5: Voiced (grey)/ whispered (orange) *reconstructed curves. Top left /s, s_W/, top right /z, z_W/, bottom left /ʃ, ʃ_W/ and bottom right /ʒ, ʒ_W/.*

3.2. FPCA of Not Whistled and Whistled PSD

When manually analysing fricatives' spectral slopes, a spectral peak was observed above the frequency of the broad peak (the first resonance of the from cavity), that could have an impact on the estimation of parameters m1 and m2 (some of these parameters' values were not significantly different in the two speech modes). This peak is likely to correspond to a whistle that was coupled into the second resonance of the front cavity (Shadle & Scully, 1995). Tuned whistles, "reinforcing" the fricative's broad peak have also been observed by Kim et al. (2014).

In this subsection, we present the results of an exploratory analysis of whisling in alveolar fricatives /s, z/, as produced by the 8 female speakers. When a spectral peak, resulting from what Pinto and Sadowsky (2019) described as an "ultra-high-frequency" whistle, was observed between 9 and 13 kHz, it was annotated manually. **Figure 6** shows two examples of whistled fricatives. This analysis revealed 166 not whistled (69 voiced and 97 whispered) and 402 whistled (219 voiced and 183 whispered) tokens of /s/, and 113 not whistled (49 voiced and 64 whispered) and 154 whistled (89 voiced and 65 whispered) tokens of /z/.

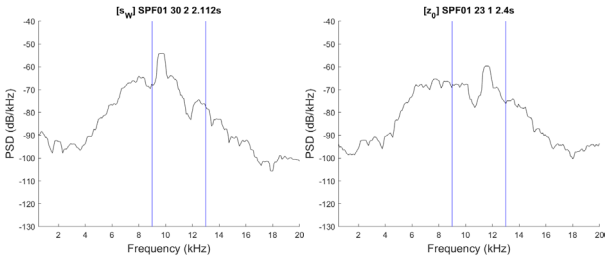


Figure 6: *Whispered /s/(left) and devoiced /z/ (right).*

We then computed the first two principal components of the not whistled/ whistled curve pairs, shown in **Figure 7** for words, which explained 93.3 % (PC1) and 6.7 % (PC2) of the /s/ variance, and 90.4 % (PC1) and 9.6 % (PC2) of the /z/ variance. Results of modelling the curves based on equations (3) and (4) revealed the following regarding the models predicting s1: /s/ – sd = 969.55, R^2 = 0.170, p = 0.000, /z/ – sd = 931.41, R^2 = 0.105, p = 0.000; and the models predicting s2: /s/ – sd = 257.32, R^2 = 0.009, p = 0.248, /z/ – sd = 299.57, R^2 = 0.051, p = 0.007.

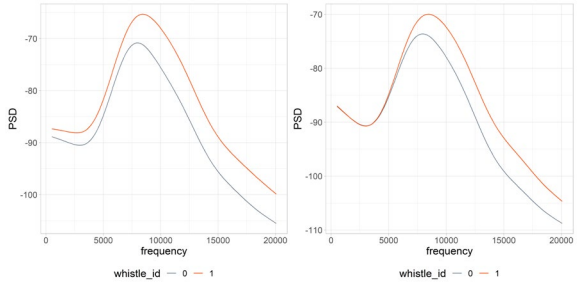


Figure 7: *Not whistled (grey)/ whistled (orange) reconstructed curves. Left /s/ and right /z/.*

3.3. Durations

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 8**) 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.

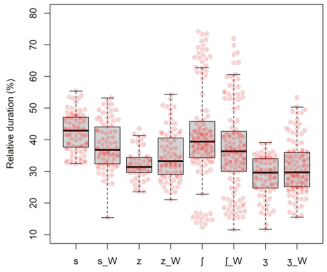


Figure 8: *Male voiced /s, z, ʃ, ʒ/ and whispered /s_W, z_W, ʃ_W, ʒ_W/ fricatives' relative duration in words.*

4. Discussion and conclusion

m1 and F_{BP} were attributes associated with consonantal place of articulation and the relative duration carried sufficiently distinct information to disambiguate consonant voicing both in voiced and whispered speech.

The fricatives' source strength (related 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 of 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.

Modelling of PSD voiced/ whispered pairs' FPCA scores revealed different sources of variation for /s, ʃ/ and /z, ʒ/.

Whistled fricatives were observed both in voiced and whispered speech, so this does not seem to be a mechanism used to compensate for a weaker source strength as typically observed in whispered speech. Nevertheless, reconstructed not whistled/ whistled curve pairs were significantly different.

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.

5. Acknowledgements

This work was financially supported by Project PTDC/ EMDEMD/ 29308/ 2017 - POCI-01-0145-FEDER-029308 - funded by FEDER funds through COMPETE2020 - Programa Operacional Competitividade e Internacionalização (POCI) and by national funds (PIDDAC) through FCT/MCTES. Support was also received from National Funds through the FCT - Foundation for Science and Technology, in the context of the projects UIDB/00127/2020 and UIDB/04106/2020.

6. References

Ball, M. J., & Local, J. (1996). Current developments in transcription. In M. J. Ball & M. Duckworth (Eds.), *Advances in Clinical Phonetics* (pp. 51–90). John Benjamins.

Benninger, M. S., Finnegan, E. M., Kraus, D. H., Sterman, B. M., Miller, R., Carwell, M. A., & Levine, H. L. (1988). The whisper and the whistle: The role in vocal trauma. *Medical Problems of Performing Arts*, *3*(4), 151–154.

Cronenberg, J., Gubian, M., Harrington, J., & Ruch, H. (2020). A dynamic model of the change from pre- to post-aspiration in Andalusian Spanish. *Journal of Phonetics*, *83*, 101016.

Gubian, M. (2023). *Modelling multi-dimensional and misaligned time-varying contours (teleconference)*. University of Konstanz, Germany, 25-26 September 2023.

Happ-Kurz, C. (2020). Object-Oriented Software for Functional Data. *Journal of Statistical Software*, *93*(5).

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.

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*.

Jovičić, S. T., & Šarić, Z. (2008). Acoustic analysis of consonants in whispered speech. *Journal of Voice*, *22*(3), 263–274.

Kim, S., Kawahara, S., & Lee, S. J. (2014). The "Whistled" Fricative in Xitsonga: Its Articulation and Acoustics. *Phonetica*, *71*(1), 50–81.

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.

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.

Pinto, L. P., & Sadowsky, S. (2019). The Ultra-High-Frequency Whistled /s/ of Southern Chilean Spanish: Socioeconomic and Gender Stratification of its Spectral Moments and Prevalence. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS 2019), Melbourne, Australia*, 48–52.

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.

Shadle, C. H. (1983). Experiments on the acoustics of whistling. *The Physics Teacher*, *21*(3), 148–154.

Shadle, C. H. (2010). The Aerodynamics of Speech. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (2nd ed., pp. 39–80). Blackwell.

Shadle, C. H., & Scully, C. (1995). An Articulatory-Acoustic-Aerodynamic Analysis of [s] in VCV Sequences. *Journal of Phonetics*, *23*(1, 2), 53–66.

Sharifzadeh, H. R., McLoughlin, I., & Russell, M. J. (2012). A comprehensive vowel space for whispered speech. *Journal of Voice*, *26*(2), e49–e56.

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.

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.

Zhang, C., & Hansen, J. (2007). Analysis and classification of speech mode: Whispered through shouted. *Proceedings of Interspeech 2007*, 2289–2292.

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.

Schwa optionality in verbal inflection in German: the effects of stress and phonetic context

Marie-Theres Weißgerber[1]

[1]*Humboldt University Berlin, Germany*
marie-theres.weissgerber@hu-berlin.de

Abstract

Speech sounds serve the function of distinguishing meaning. However, in German, schwa sometimes has no such semantic function and is optional in certain cases, like first-person singular inflectional suffixes. Nonetheless, this optionality has not yet led to a total removal, or obligation to articulate, schwa in these suffixes. The present study investigates the effects of phonetic context and stress on schwa optionality. The data set consists of two registers, formal and informal, of German spoken in Germany and Namibian German. The following speech sound and the stress of the following syllable, which are thought to affect the likelihood of word-final schwa production, were analysed for 44 speakers using Praat. Significant effects were found for stress, with less schwa productions before unstressed syllables. Significantly less schwa instances were observed before vowels. Overall, register had a significant effect, with more schwa productions in the formal condition. The impact of stress and the effect of register were more marked in the subset of Namibian German. These findings highlight the importance of investigating the interaction of phonetic features and register and emphasise the value of exploring different varieties in the study of speech production phenomena.

Keywords: spontaneous speech, phonetic context, stress patterns, register, Namibian German

1. Introduction

Schwa is optional in German first-person singular verbal inflectional suffixes. Variation in schwa realisations has been documented as a constant feature of the German language system for centuries (Fleischer et al. 2018; Nübling et al. 2013; Eisenberg 2020). The use of schwa entered the phonemic system during the Old High German period[1]. Old High German word-final unstressed vowels were the full vowels <a>, <e>, <i>, <o>, <u> (Fleischer et al. 2018). Example (1) demonstrates the shift from full vowels in the final syllable to the inflection with schwa in New High German.

(1) OHG *suochu* 'I search' (1. ps. sg. ind. pres.), *suochi* 'search' (imp. sg.), *suoche* 'may he search' (3. ps. sg. subj. pres.) > MHG *suoche* > NHG *such(e)*
(Fleischer et al. 2018)

As can be deduced from example (1), schwa in inflection has become markedly more flexible in present-day German. In some cases, like in inflectional paradigms used to form the past tense without ablaut, a schwa suffix is obligatory. For instance, the verb "sehen" ('to see') forms the third-person singular with "sieht" in the present tense and with "sah" in the preterite. In other cases, word-final schwa can be either pronounced or omitted without yielding any semantic change. Figure 1 shows verbal inflection for a weak verb with schwa in the first-person singular in present tense as an optional suffix. The third-person singular in present tense and the stem of the first-person singular in preterite tense are identical on the surface level. Schwa is not optional in the first and third-person singular in preterite tense. If schwa was not realised in these cases, the result would be a semantic change of the verb form towards the present tense.

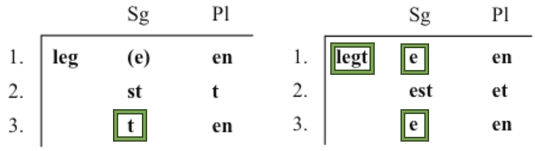


Figure 1: Inflection of the weak verb "legen" ('to put') in present tense (left side) and preterite (right side), modified with green boxes by author (Eisenberg 2020).

Schwa 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. (2018) optionality in word-final schwas is examined in adverbs. 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. 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. 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). On the supra-segmental level, word stress plays a crucial role in triggering the presence or absence of word-final schwa. Research that considers stress patterns and their influence on schwa alternations often uses the

[1]The period of Old High German with attested writing is dated to 750-1050 AD. The Middle High German (MHG) epoch is dated to 1050-1350, and the time period of New High German (NHG) began in 1650 and continues to this day (Nübling et al. 2013).

terms ‘stress clash’ and ‘stress lapse’. For example, Kentner associates ‘rhythmic alternation’ with the avoidance of ‘stress clash’ and ‘stress lapse’ (Kentner et al. 2018). Another crucial concept is the tendency of German rhythm to adhere to a pattern of sequential trochees, which is referred to “as an optimal template regulating the shape of words” in German (Kentner et al. 2018, p. 120). However, it is not only within words that the trochee plays an important role. On the sentence level, a balanced juxtaposition of trochees prevents a “clustering of strong syllables (*CLASH)” as well as “sequences of weak syllables (*LAPSE)” (Kentner et al. 2018, p. 120). Within the framework of *Prosodic Parallelism Theory*, Wiese and Speyer (2015) ascribe an important role to schwa. As part of this framework, the authors argue that “a form that is invariably (non-)trochaic causes another form in the same dominating category[2] (the phrase) to be (non-)trochaic as well.” (R. Wiese and Speyer 2015, p. 528). While the framework allows for exceptions where lexical options are limited (“as in *sehr langsam* ‘very slow”’), the authors assume that prosodic parallelism is the favoured choice wherever feasible. They claim that word-final schwa optionality offers the opportunity for such a selection.

Yet, contributing factors for schwa-zero alternations are not only found on a purely linguistic level – whether or not schwa is articulated also seems to depend on situational and task-based factors. 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 and 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 are 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 (H. 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 (H. Wiese and Bracke 2021).

This study investigates how schwa in the first-person singular is distributed in spontaneous speech in two registers, formal and informal, in two varieties of German. 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.

[2]This assumption is based on a hierarchy of phrase, word, foot and syllable (in descending order) (R. Wiese and Speyer 2015).

2. Methods

2.1. Corpora

Speech recordings were obtained from two corpora containig spontaneous speech recordings. Data of native speakers of German residing in Germany stem from the RUEG corpus (H. Wiese, Alexiadou, et al. 2021). Recordings of native speakers of Namibian German have been made available by the research group *Namdeutsch* (‘Namibian German’) within the scope of the corpus *DNam (Deutsch in Namibia)* (‘German in Namibia’) (Zimmer et al. 2020).

2.2. Participants and Tasks

Participants were presented with visual material, either in the form of a video or a photograph story, of an accident. After viewing the material, speakers 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, participants summarised events to a friend in a voice message (H. Wiese, Alexiadou, et al. 2021; H. 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, speakers 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. The age range of the speakers is between 13 and 40.

2.3. Stimuli and Analysis

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) on five different tiers (see Figure 2). Tier one contains transcriptions of the spoken materials. Tier two contains the relevant stimulus with first-person singular inflection. Tier three shows whether or not schwa is articulated in the relevant stimulus. Here, *1* refers to a realisation of schwa and *0* denotes a non-realisation. In tier four, the phonological context preceding and following the expected schwa realisation is described. In tier five the stress pattern of the following syllable is annotated. Factors influencing schwa realisation were tested using chi-square tests.

3. Results

3.1. Phonetic context

To assess the effect of phonetic context, the data were subset into instances of verbs preceding vowels, sonorants, obstruents and pauses (see Figure 3). As can be seen, a following pause, indicating a prosodic boundary, increases the frequency

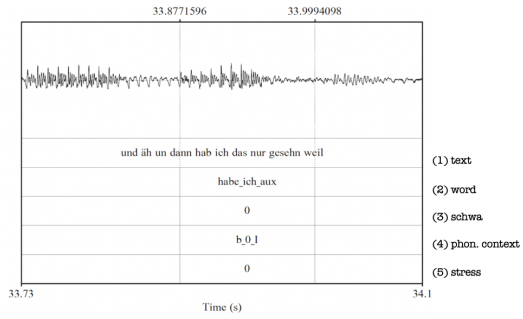


Figure 2: Textgrid example.

of realised schwas. However, only 14 instances with following pauses were found. The subset of unrealised schwas preceding vowels are distributed to 94.3% before unstressed vowels, non-realised schwa before sonorants can be found to 92.3% before unstressed sonorants, and the subset of unrealised schwas before obstruents are distributed to 82.8% before unstressed obstruents. This result indicates a slight sonority continuum within the distribution of schwa realisations and their interaction with the following context.

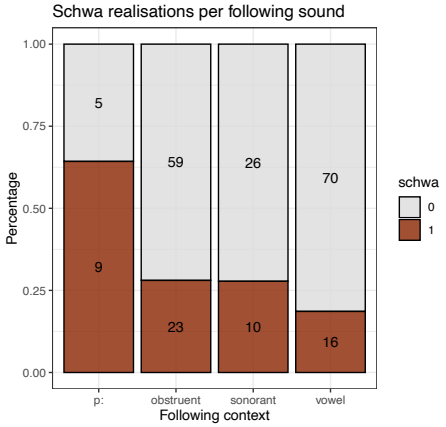


Figure 3: Schwa-realisations coded by following segment, absolute numbers in bars.

3.2. Stress

Out of all instances, merely 32 (14.7%) are followed by a stressed syllable, 172 (78,9%) precede an unstressed syllable, and 14 (6.4%) are followed by a pause. Excluding items with following pauses, a Pearson's Chi-squared test 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. In the GGER data frame, 60% of potential word-final schwas are articulated before stressed syllables. This is the case in only 45.5% in the NamGER subset (see Figure 4).

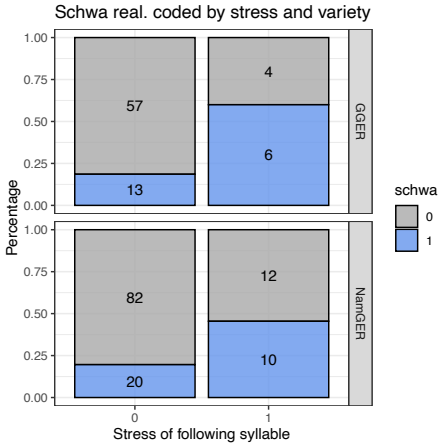


Figure 4: Schwa realisations per variety coded by the stress pattern of the following syllable, absolute numbers in bars. Pauses are excluded.

3.3. Register

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 ($\chi^2 = 15.009$, $df = 1$, $p < .001$). In the formal register, NamGER verbs are pronounced with schwa in 63.4%. The proportions are different between the two varieties in the informal condition. As can be seen in Figure 5, there is a difference in the informal condition: in NamGER 10% of the schwas are realised, whereas in GGER it is 15%.

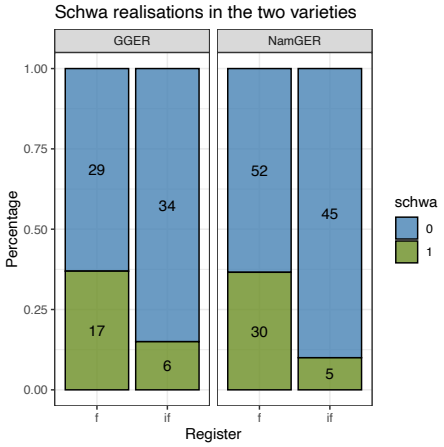


Figure 5: Schwa realisations per variety coded by register, absolute numbers in bars.

4. Discussion and conclusion

In summary, the results demonstrate that the most common realisation of the verbal inflectional ending in the first-person singular is without schwa in 73.4% of all cases ($n = 160$), similar to Lange et al. (2023, accepted). Based on the literature (Fleischer et al. 2018; Kohler and Rodgers 2001), it was expected that schwa should be realised less often when it precedes a vowel. This can be confirmed with the data set analysed in the present

study, where the effect of following vowels on schwa realisations is statistically significant. The stress of the following segment has a significant influence on whether or not a schwa is realised. In general, the realisation of schwa is significantly influenced by the stress of the following syllable. However, there appears to be no interaction between stress and register. The fact that adjacent rhythmic context affects schwa realisations supports the findings of Wiese and Speyer and Kentner (R. Wiese and Speyer 2015; Kentner et al. 2018).

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. (H. 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?

Schwa is indeed optional in first-person singular verbal inflectional suffixes in German. This study provides further evidence demonstrating that this optionality is not random. In fact, schwa optionality is found in both the formal and the informal register. The presence of optionality in the informal register highlights that register has an effect on schwa realisations in the German language. Variation in the formal register, on the other hand, demonstrates that this variation is not confined to informal communicative settings – schwa optionality seems to be ingrained in first-person singular verbal inflectional suffixes of German. Additionally, this study offers insights into how this phenomenon operates in Namibian German. The results indicate that, given that particular elements are even more pronounced in this variety, Namibian German might be advancing an internal structural phenomenon of the linguistic system of German. This study confirms previous findings on the interaction of phonetic context and stress with word-final schwa. It will be a worthwhile endeavour to further examine which individual elements cause these marked differences in schwa realisations by investigating other registers and varieties.

5. Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334. The author would like to thank Prof. Christine Mooshammer, Prof. Heike Wiese and Dr. Malte Belz for guidance and feedback as well as Robert Lange, Daniela Palleschi and Georg Lohfink for technical support.

6. References

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.

Kentner, Gerrit, Christiane Ulbrich, Alexander Werth, and Richard Wiese (2018). "Schwa-optionality and the prosodic shape of words and phrases". In: *Empirical approaches to the phonological structure of words*, pp. 121–151.

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.

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, Shanley Allen, Oliver Bunk, Natalia Gagarina, Kateryna Iefremenko, Jahns Esther, Martin Klotz, Thomas Krause, Annika Labrenz, Anke Lüdeling, Maria Martynova, Katrin Neuhaus, Tatiana Pashkova, Vicky Rizou, Rosemarie Tracy, Christoph Schroeder, Luka Szucsich, Wintai Tsehaye, Sabine Zerbian, and Yulia Zuban (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.

Wiese, Richard and Augustin Speyer (2015). "Prosodic parallelism explaining morphophonological variation in German". In: *Linguistics* 53.3, pp. 525–559.

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.

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

Abstract

The merger of the Mandarin [s]~[ʂ] contrast, known as "deretroflexion", frequently occurs in Mandarin spoken by bilingual Southern Min speakers, whose L1 lacks the retroflex category. This study explores the production of the Mandarin alveolar-retroflex contrast by bilingual speakers of Quanzhou Southern Min (L1) and Mandarin (L2) in two different vowel contexts ([a] vs. [u]). Our bilingual speakers' contrast production was evaluated using a perceptual identification task by L1 Mandarin speakers, showing only a small subset of our sample who maintained the [s]~[ʂ] contrast. We found significant Center of Gravity (CoG) differences between the two target fricatives for "distinctive" speakers, with this difference being larger in the context of [a] than [u]. For all speakers, the acoustic difference between the target fricatives increased with increased exposure to and use of Mandarin.

Keywords: sibilant fricatives, contrast merger, Mandarin, Quanzhou Southern Min

1. 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". This process, detailed in Kubler (1985), underscores how language contact with L1 Southern Min, which lacks the retroflex phone, led to a notable convergence of the retroflex sibilants towards an alveolar pronunciation in Mandarin. This reflects a broader pattern of L2 phonological adaptation in response to the phonological inventories of the languages in contact.

Other linguistic factors, such as vowel context, have also been noted to influence this contrast merger, but some conflicting results have emerged. On the one hand, Chang and Shih (2015) demonstrated a notable influence of vowel context on the spectral differentiation between alveolar and retroflex fricatives in both Beijing Mandarin and bilingual speakers of Mandarin and Taiwan Southern Min. In comparison to the [a] vowel context, it was observed that, in the [u] context, speakers from both regions exhibited a reduced spectral contrast. On the other hand, Chiu et al. (2020) applied ultrasound imaging techniques to the variablility of sibilant contrast production, and found that the tongue postures for [s] and [ʂ] showed more "context-dependent overlap" in the context of [a].

The exploration of variability in the merger of retroflex and alveolar sibilants extends, however, beyond purely linguistic dimensions. Recent research suggests that production variability in the merger of this sibilant contrast can additionally be captured by considering social factors, such as age, gender, and language exposure level (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 (泉州闽南话, henceforth QSM) [L1] and Mandarin [L2] and thus examines different linguistic and social factors at both the group and individual levels.

2. Method

61 bilingual speakers of QSM and Mandarin (29 men, 32 women) were recruited in Quanzhou, China, divided into three age ranges between 18 and 55 (18–30: 27 participants, 31–40: 18 participants, 41–55: 16 participants). These participants all have self-reported native-level fluency in Quanzhou Southern Min and Mandarin. They had all spent their childhood in Quanzhou and were living there at the time of the study. Mandarin was used as a metalanguage in experimental materials (including on-screen instructions), but all communication with the experimenter before, during, and after experimental sessions was conducted in Quanzhou Southern Min.

Each participant took part in a sentence reading task with target words embedded into carrier sentences, e.g., "请阅读单词X八遍", "Please read the word X eight times". Targets were all real Mandarin words of the form CVCV (2 fricatives × 2 vowel contexts ([a] vs. [u]) × 3 examples) and realized with a high level tone (tone 1) on the first syllable ([i] was not included as neither [si] nor [ʂi] are phonotactically well-formed in Mandarin). Target words were all represented orthographically as two Simplified Chinese characters. The lexical frequency of each real word was controlled to be within the log frequency range of 3 to 5 according to the SUBTLEX-CH corpus (Cai and Brysbaert 2010). Recordings were made in a quiet room at a sampling rate of 44.1 kHz using a Neumann TLM102 microphone, and a USBPre 2 audio mixer by Sound Devices. To guarantee the quality of the recording, we placed Alctron's VB 860 noise-canceling filter around the recording setup and installed soundproofing foam on both the window and the door of the room. We also ensured that the noise levels were maintained below -48 dB with the help of a Benetech GM1356 Digital Sound Decibel Noise Level Meter Tester. Before conducting the experiment, ethical approval was obtained at the Université Paris Cité (IRB Number: 00012022–95).

3. Results

Data from one participant were excluded because he had difficulty reading the Chinese characters during the task. Data ana-

lyzed are from the remaining 60 participants. Spectral moments were extracted at the mid-point of each fricative using the Praat script of DiCanio (2013). We focus here on Center of Gravity (CoG). We first compared speakers' [s] and [ʂ] productions in the two vowel contexts, shown in fig. 1. We employed mixed-effects models to investigate the effects of Contrast and Vowel, as well as their interaction, on CoG values while accounting for individual variability with by-participant random intercepts. We compared this full model to reduced models using likelihood ratio tests and found that the full model was a significantly better fit to the data than models which excluded the factors Fricative ($\beta = -343.2$, $SE = 48.02$, $\chi^2(1) = 49.40$, $p < 0.001$) and Vowel ($\beta = -548.28$, $SE = 48.02$, $\chi^2(1) = 119.47$, $p < 0.001$). The full model did not significantly differ from a reduced model which excluded the interaction between the two factors ($\chi^2(1) = 3.05$, $p > 0.05$). While the statistical model underscores significant contrast at the group level, along with a general coarticulatory effect (lower CoG values in the context of [u]), the graphical representation shows considerable overlap among the data points. We therefore sought to measure how individual participants produced the target fricatives, in order to categorize individuals as producing a reliable contrast or merging the two fricatives.

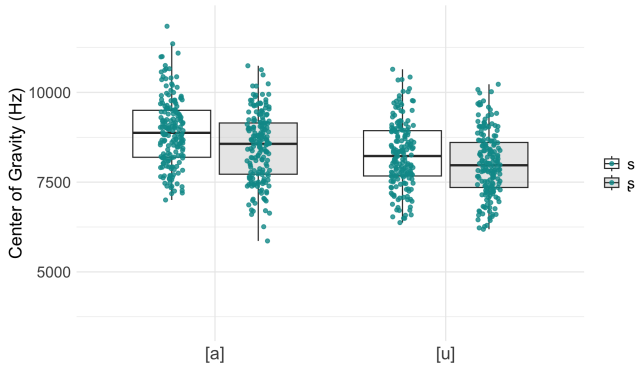


Figure 1: Comparison of CoG value for [s] and [ʂ] fricatives across vowel contexts within bilingual QSM Speakers

To investigate individual-level variability, we conducted a perceptual coding study involving ten native Mandarin speakers (5 men and 5 women, with a mean age of 28 years). These participants were recruited to perform a two-alternative forced-choice identification task. The aim was to assess the judgments of native Mandarin speakers regarding the productions of our bilingual QSM speakers. Stimuli consisted of CV syllables extracted from the sentence reading task performed by the 60 QSM speakers (12 tokens × 60 speakers). For each trial, the L1 Mandarin listeners heard a token of one of the QSM speakers' productions and had to indicate if they thought it corresponded to [s] or [ʂ]. Participants saw two Simplified Chinese characters and corresponding pinyin which indicated the response options, for example "sū苏" or "shū书".

We computed the L1 Mandarin speakers' identification overall accuracy for each QSM speaker in both [a] and [u] contexts. Figure 2 and fig. 3 summarize how the native listeners identified the sibilants produced by individual QSM speakers. The x-axis represents identification accuracy of individuals' [ʂ]-targets and the y-axis represents identification accuracy of individuals' [s]-targets. Following Chang and Shih (2015), we consider reliable productions to be above the threshold of 60% identification accuracy. Individuals had to produce both [s]- and [ʂ]-targets above this accuracy threshold in order to be considered to make a reliable sibilant contrast; they are shown inside the box in the top right corner of the figures. Consequently, QSM participants such as speaker 12, who demonstrated accuracy rates above 60% for both [s] and [ʂ] are classified as "distinctive" speakers. In contrast, many participants fall into the "merged" category due to their significantly lower accuracy rate (below 60%) for both fricatives. Among those classified as "merged", variability in contrast accuracy persists. For example, speaker 7 was classified as "merged" due to the significantly lower identification accuracy rate of their [ʂ]-targets (close to zero), despite an [s]-target accuracy nearly reaching 100%. Such speakers are producing fricatives that are perceived by L1 Mandarin speakers as [s] across the board (yielding high accuracy for [s]-targets and near-zero accuracy for [ʂ]-targets). These speakers are clustered in the top-left of the figures. On the other hand, speaker 22's [ʂ]-target production (in the bottom-right of the figures) achieves close to 100% accuracy but this speaker's [s]-targets were identified with near-zero accuracy. This speaker is producing fricatives that are perceived by L1 Mandarin speakers as [ʂ] across the board, a likely case of hypercorrection. Other speakers fall somewhere between these two extremes, producing some fricatives that are accurately perceived by L1 Mandarin speakers, but not above the 60% threshold.

We identified 9 QSM speakers who produced a reliably "distinctive" contrast in both vowel contexts, in contrast to 48 speakers who were categorized as "merged" in both vowel contexts. There were three additional participants who demonstrated the ability to distinguish the target contrast in one vowel context but not the other ([a]: speakers 15 and 30; [u]: speaker 5). For the sake of brevity, we focus in the rest of the paper on the 57 distinctive and merged participants.

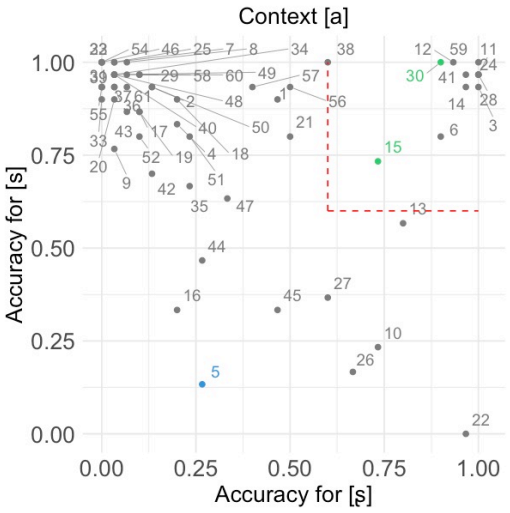


Figure 2: L1 Mandarin speakers' identification accuracy of QSM bilinguals' [s]- and [ʂ]-target productions in the context of [a].

3.1. Linguistic effects

CoG values for "distinctive" and "merged" speakers are shown in fig. 4. For both "distinctive" and "merged" speakers, we

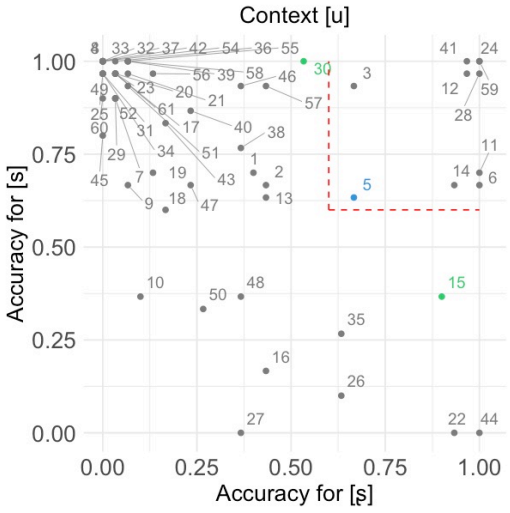


Figure 3: L1 Mandarin speakers' identification accuracy of QSM bilinguals' [s]- and [ʂ]-target productions in the context of [u].

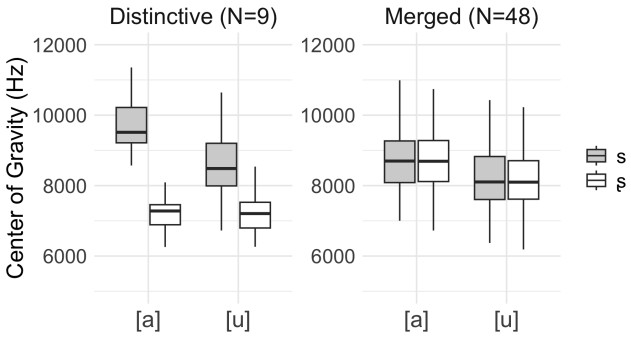


Figure 4: Comparison of CoG value for [s] and [ʂ] fricatives across vowel contexts between "distinctive" and "merged" bilingual QSM speakers

employed mixed-effects models to investigate the effects of Fricative and Vowel (both included using deviation coding), as well as their interaction (Fricative × Vowel), as fixed factors on CoG values, while accounting for individual variability of QSM speakers 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 "distinctive" speakers, the full model was a significantly better fit to the data than models which excluded the factors Fricative ($\beta = -1978.2$, $SE = 127.9$, $\chi^2(1) = 123.8$, $p < 0.001$), Vowel ($\beta = -531.4$, $SE = 127.9$, $\chi^2(1) = 16.3$, $p < 0.001$), and their interaction ($\beta = 1082.4$, $SE = 255.8$, $\chi^2(1) = 16.9$, $p < 0.001$). This finding confirms that for these speakers who were perceived as producing different [s]- and [ʂ]-targets, their CoG values significantly differed according to the target fricative. Additionally, alongside the previously reported general coarticulatory effect (lower CoG values before [u]), the "distinctive" speakers exhibited a greater difference in the CoG values between alveolar and retroflex fricatives in the context of [a] than in the context of [u]. This suggests that "distinctive" speakers are able to maintain a greater spectral contrast between [s] and [ʂ] when followed by [a]. This observation aligns with the research presented by Chang and Shih (2015) (cf. Chiu et al. 2020), which noted that speakers displayed a larger spectral contrast distance in the [a] context compared to the [u] context, the rounded vowel tending to reduce the CoG in the realization of alveolar and retroflex fricatives.

For "merged" speakers, the analysis revealed that only the factor Vowel significantly affected model fit ($\beta = -544.8$, $SE = 40.2$, $\chi^2(1) = 158.3$, $p < 0.001$). This significant effect underscores again that the CoG values for [s] and [ʂ] are affected by the vocalic context, with both showing higher CoG values in the context of [a] compared to [u]. However, the factor Fricative does not exert a significant effect on the model fit for "merged" speakers ($\chi^2(1) < 1$). Similarly, the interaction between Vowel and Fricative does not contribute significantly to the model fit ($\chi^2(1) < 1$). These results suggest that, for "merged" speakers, the contrast between [s] and [ʂ] is not reliably maintained in production and that the fricatives that are being produced are all similarly affected by vowel context.

3.2. Extra-linguistic effects

As mentioned in the introduction, extra-linguistic (social) factors might also influence the variation we observed. We examined whether exposure to and use of Mandarin, age group, and gender influenced productions of the target fricatives. Given that our QSM participants all self-identified as highly bilingual, we focus on their L2 usage frequency. For assessing the extent of Mandarin exposure and use, we based on their responses to our post-test language use questionnaire. We followed Weng, Chitoran, and Martin (2023), which involved assigning an overall Mandarin exposure and use score to each participant. This score, which ranged from -8 to 8, was based on self-reported frequency of use of Mandarin and QSM on a five-point scale from "always QSM, never Mandarin" (-2), to "half QSM/half Mandarin" (0) to "always Mandarin, never QSM" (2) across four contextual domains: language used in childhood, within family settings, with friends, and among colleagues. A higher score indicates greater and more consistent exposure to and use of Mandarin relative to QSM. We observed variation in participants' responses ($M = -0.46$, $SD = 2.45$; recall that a score of 0 represents balanced Mandarin/QSM usage).

Because CoG values were found to significantly differ according to vowel context for both "distinctive" and "merged" speakers, we looked at data from each vowel context separately (see fig. 5). For each vowel context, we created a linear regression model to predict the average CoG differences of the participants in that context (each participant's average [s]-target CoG − their average [ʂ]-target CoG). The predictors included individual Mandarin exposure scores, gender, age group, and speaker classification (distinctive vs. merged), as well as the interaction between each speaker's classification and Mandarin exposure level score. Our analysis revealed that, for both vowel contexts, speakers with a higher Mandarin exposure score tended to produce a larger contrast difference between the target fricatives ([a] context: $\beta = 176.0$, $SE = 68.2$, $t = 2.5$, $p < 0.05$; [u] context: $\beta = 216.0$, $SE = 51.5$, $p < 0.001$). Moreover, both models indicated a significant negative effect of being classified as a merged speaker ([a] context: $\beta = -2435.5$, $SE = 193.6$, $t = -12.5$, $p < 0.001$; [u] context: $\beta = -1301.9$, $SE = 146.4$, $t = -8.8$, $p < 0.001$), again reflecting that distinctive speakers maintained a larger

CoG difference between the target fricatives, showing a clear acoustic contrast. It appears that despite observing an increase in contrast CoG difference between distinctive and merged speakers as Mandarin exposure score rise, the interaction between a speaker's classification and their Mandarin exposure score does not show a significant effect on CoG differences for either [a] or [u] context ([a] context: $\beta = -157.7$, $SE = 82.3$, $t = -1.9$, $p = 0.06$; [u] context: $\beta = -109.3$, $SE = 62.3$, $t = -1.7$, $p = 0.08$). The limited sample size of nine data points for the distinctive group in each vowel context may be a contributing factor to this outcome. Such a small dataset can limit the statistical power of the study, potentially obscuring real effects that might emerge with a larger number of observations. Consequently, while increased Mandarin exposure seems to be associated with the production of larger CoG differences, the current evidence does not conclusively support a differential impact based on speaker classification.

Concerning the other social factors (age group, gender), for the [a] context, speakers in the middle age range (41–55) showed a significant difference with the youngest group ($\beta = -434.46$, $p < 0.01$; all others $p > 0.05$). In the context of [u], significant effects were observed for gender and age, with women [compared to men] ($\beta = -283.98$, $p < 0.05$) and younger speakers [compared to the middle and older groups] producing more distinct contrasts (young vs. middle: $\beta = -471.88$, $p < 0.001$; young vs. older: $\beta = -423.93$, $p < 0.01$).

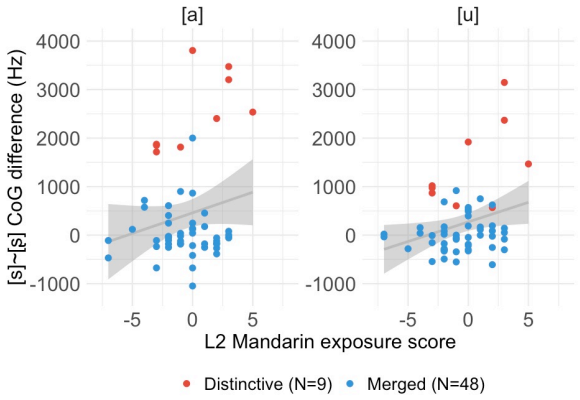


Figure 5: Participants' mean CoG difference as predicted by L2 Mandarin exposure level in each vowel context. More positive scores represent higher exposure to and use of Mandarin compared to QSM; more negative scores represent higher exposure to and use of QSM compared to Mandarin.

4. Discussion

In this study, we tested the production of a Mandarin sibilant fricative contrast by bilingual speakers of Quanzhou Southern Min (L1) and Mandarin (L2) in two different vowel contexts. Our results indicate that both the following vowel and a speaker's Mandarin exposure level are significant predictors of how this contrast is produced. Through the perception judgments of L1 Mandarin speakers, we categorized our bilingual speakers into two distinct groups: "distinctive" and "merged". Both groups showed a coarticulatory effect such that CoG values of each fricative were lower before the vowel [u]. Meanwhile, the 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 larger CoG difference between the target fricatives). However, a significant interaction effect was not observed for either group. We speculate that this may be due to the disparity in sample sizes, with only 9 "distinctive" speakers compared to 48 who were categorized as "merged". This imbalance could potentially alter the interpretation of interaction effects. Further research with more "distinctive" speakers is needed to make these findings clearer and see if the trend we noticed (with a stronger effect for distinctive as compared to "merged" speakers) holds true.

Our analysis also identified patterns of hypercorrection and hypocorrection among the "merged" speakers' productions, suggesting a variety of profiles. This raises the question: 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. Additionally, it is yet to be determined if speakers who merge contrasts in production also do so in their perception, highlighting a potential area for future research to explore the relationship between production and perception in bilingual populations.

5. Acknowledgements

Funding for this study was provided through a doctoral contract with Université Paris Cité, and through partial funding from the ANR-10-LABX-0083-LabEx EFL. The authors declare no conflicts of interest.

6. 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.

Chiu, Chenhao, Po-Chun Wei, Masaki Noguchi, and Noriko Yamane (2020). "Sibilant fricative merging in Taiwan Mandarin: An investigation of tongue postures using ultrasound imaging". In: *Language and speech* 63.4, pp. 877–897.

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*.

DiCanio, Christian (2013). *Spectral moments of fricative spectra script in Praat*. Haskins Laboratories & SUNY Buffalo.

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.

Weng, Caihong, Ioana Chitoran, and Alexander Martin (2023). "Bilingual phonological contrast perception: The influence of Quanzhou Southern Min on Mandarin non-sibilant fricative discrimination". In: *JASA Express Letters* 3.7.

An exploration of pitch in Afro-Mexican Spanish

Gilly Marchini[1]

[1]*University of Edinburgh*

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

Abstract

This paper presents a descriptive analysis of pitch in Afro-Mexican Spanish, a largely unexplored variety of Spanish spoken by mixed indigenous-African communities in Southwestern Mexico. Sociolinguistic interview data was collected from one female speaker (51 years old), with a total of 122 broad focus, declarative Intonational Phrases annotated according to Sp_ToBI protocol.

Results reveal that whilst Afro-Hispanic language employs pitch at a word-level, i.e., as cue to lexical stress, pitch is phrasal in Afro-Mexican Spanish, thus aligning with non-Afro Mexican varieties: broad focus declaratives are signalled through their low tone pre-nuclear pitch accents and circumflex nuclear accents. However, whilst peaks align on the stressed syllable across open and closed syllables, there is an interaction with the nasality of the following sound: if present on the segmental string, peaks align on the following, post-vocalic nasals (/n, m, ŋ, ɲ/ in the current dataset) regardless of intervening syllable boundaries. In the case of closed syllables, i.e., with coda /N/, e.g., descendiente [de.sen.ˈdjen.te] (‘descendent’) , peaks align tonically (90.5% of instances). For open syllables, i.e., with /N/ as following onset, e.g., mexicano [me.xi.ˈka.no] (‘mexican’), peaks align post-tonically (100% of instances).

Although tonic peak alignment is common across Afro-Hispanic varieties, the role of the nasal is unexpected. Nor is it common in non-Afro Mexican Spanishes, where instead peaks are often displaced, reaching their maxima on the following, post-tonic syllable. Despite established theoretical claims that peak alignment should not vary according to segments (The Segmental Anchoring Hypothesis, or SAH), these preliminary results indicate that this does not occur for this dialect. I therefore consider the bearing of this upon the dialect-specific nature of the SAH, with reference to control experimentation required to test whether such features are unique to the dialect.

Keywords: speech production, speech synthesis

1. Introduction

This paper documents the features of pitch in Afro-Mexican Spanish, an under-researched variety of Spanish spoken by 37 communities of mixed indigenous and African heritage on the Costa Chica, Mexico (Guerrero/Oaxaca). Background on Afro-Hispanic and non-Afro Spanish prosody is provided in §2, with the methods outlined in §3. I first analyse the variation in pre-nuclear pitch accent, and the distribution of tonically and post-tonically aligned peaks in §4.1, prior to the variation of nuclear accent configuration in §4.2. §5 discusses how findings diverge from those described in non-Afro varieties, and their theoretical bearing on pitch anchoring processes.

2. Background

Afro-Hispanic language is an umbrella term for Spanishes spoken by those of African heritage. With the exception of Palenquero (Hualde and Schwegler 2007), it is accepted that these are not creoles but varieties of Spanish with a number of shared features (Arends and Bruyn 1994; Lipski 2010; Sessarego 2015; Schwegler 1999; Schwegler 2001). Of particular interest to the current paper are the use of pitch as a word-level correlate and differences in peak alignment.

2.1. Pitch

Within an autosegmental-metrical framework, stressed syllables act as anchors upon which pitch accents may dock. Generally, pre-nuclear accents are considered any pre-final accent, and nuclear any final (in combination with the final boundary tone) (Ladd 2008). In non-Afro Mexican Spanishes, pitch is a phrase-level correlate, employed as part of intonation to signal differences in information structure. In this way, the interrogative *¿Vas a la tienda?* (‘You are going to the shops?’) and the declarative *Vas a la tienda.* (‘You are going to the shops.’) are distinguished through the global high rising pitch in the former and global low-falling pitch in the latter. Specific to broad focus declaratives, i.e., utterances in which no one element is emphasised as more important than the others, pre-nuclear pitch accents are consistently realised a low tones, or lower relative to that which has come before, and the nuclear accent as a circumflex during which F0 rises during the final tonic syllable of the utterance and then sharply falls (Mota et al. 2011; Martín Butragueño 2003; Martín Butragueño 2004; Martín Butragueño 2006; Martín Butragueño 2019; Willis 2005). Although the velocity of the fall and the rate at which the latter accent occurs varies regionally (Martín Butragueño 2004; Prieto, Shih, and Nibert 1996; Martín Butragueño 2019), these features together act a salient cue to broad focus, declaratives across non-Afro Mexican varieties. Outwith the Mexican context, research shows an interaction with utterance length: in Peninsular Spanish, the circumflex accent is more likely in utterances of 2 phonological words (p-words) than in those of 1 (Torreira and Grice 2018). Whether such effects emerge in Afro- and non-Afro Mexican varieties is unclear.

Regardless, these features are in contrast to Afro-Hispanic language where instead pitch occurs at a word-level: it is a correlate of syllable stress, such that stressed syllable are invariantly produced with a high rising pitch movement, and unstressed low. In this way, pitch acts as cue to lexical meaning, e.g.,, the distinction between penultimate stress in *animo* /a.ˈni.mo/ (‘I encourage’) and ultimate stress in *animó* /a.ni.ˈmo/ (‘He encouraged’) (Hualde and Schwegler 2007; Lipski 2004; Lipski 2006; Lipski 2008; Sessarego 2015). Whether pitch conveys phrase- or word-level meaning in Afro-Mexican Spanish is a central question of this paper.

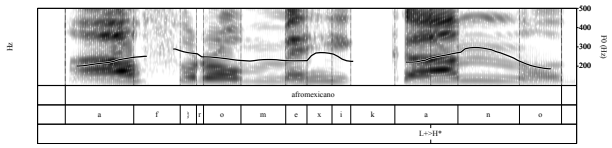


Figure 1: Pre-nuclear, post-tonic peak alignment in open syllable in 'afro-mexicano' ('afro-mexican').

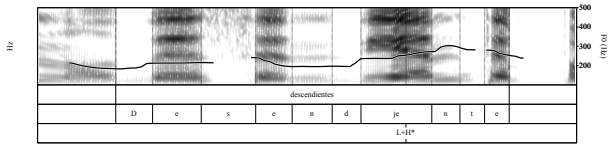


Figure 2: Pre-nuclear, tonic peak alignment in closed syllable in 'descendiente' ('descendent').

2.2. Peak alignment

Rises may be defined as pitch movements initiating from a low point (a valley) and contiguously rising to a relative high point or maxima (a peak); the L valley is thus the rise's onset, and the H peak its offset. Whilst the onset consistently docks onto the stressed syllable, peak alignment is variable (Atterer and Ladd 2004; Ladd et al. 2009; Prieto, Shih, and Nibert 1996; Prieto and Torreira 2007). In non-Afro Mexican Spanishes, pre-nuclear peak displacement is noted in broad focus declaratives: the peak is reached in the following, post-tonic syllable (Martín Butragueño 2006; Willis 2003; Willis 2005). Such displacement is not noted in Afro Hispanic varieties, where instead peaks align tonically regardless of focus or nuclear position (Hualde and Schwegler 2007; Lipski 2008; Sessarego 2015).

Impressionistically, the Afro-Mexican Spanish data reveals variability in peak alignment: although tonic peak alignment is common, it appears more likely in closed syllables than open. It is also observed that should the sound following the tonic syllable be a nasal, this is where peaks align regardless of intervening syllable boundaries. These differences are exemplified in Figures 1 and 2: in Figure 1 with /N/ as the following onset, the peak is post-tonic whilst in Figure 2 with coda /N/, the peak is tonic. Such behaviours are observational at this stage, however given their absence from both Afro-Hispanic language and surrounding non-Afro varieties, they pose interesting questions concerning pitch anchoring in this variety.

2.3. Research questions & motivation

Thus, whilst divergent prosodic features are noted in Afro-Hispanic varieties, whether these also emerge in Afro-Mexican Spanish remains unexplored. The goals of this paper are therefore two-fold. Firstly, it seeks to establish whether pitch is employed at a phrase- or word-level in this dialect. Should pre-nuclear pitch accents be invariantly high and the circumflex accent absent in nuclear position, patterns would thus align with word-level pitch in Afro-Hispanic language. Nonetheless, should pre-nuclear accents instead be generally low in their realisation, with the nuclear accent consistently realised as the circumflex, pitch may be phrase-level. Secondly, it seeks to confirm the prior observations concerning peak alignment, namely

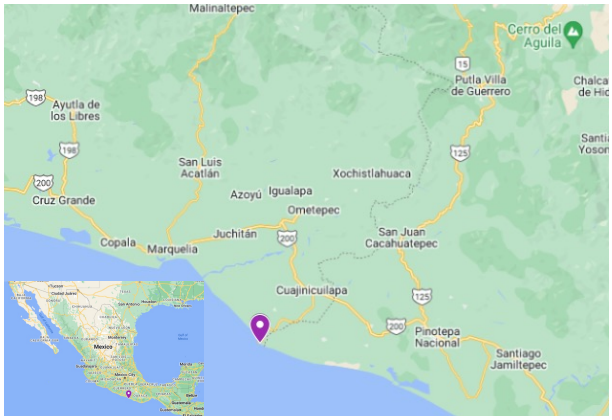


Figure 3: Map of Punta Maldonado, with location within Mexico subset in the left corner.

whether post-vocalic nasals may act as segment anchors, and the bearing of this upon future experimentation and theory. The research questions therefore ask:

RQ1. What is the variation in pre-nuclear 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?

3. Methods

Data was collected through sociolinguistic interviews recorded during a 1-month long fieldwork trip. Interviews were conducted in a group setting, with a community liaison present at all times. Discussions focused on regional history and culture. 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) (see Figure 3). 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 in order to minimise the effect of overlapping speech. Data was segmented using the MAUS aligner (Schiel 1999) and manually corrected.

Phrases were annotated according to Sp_ToBI protocol specific to Mexican varieties (Mota et al. 2011) (see Figures 1 and 2 for example). ToBI labels were extracted via Praat script. Here within, L+H* denotes tonic peaks, i.e., those reached within the stressed syllable, and L+>H* post-tonic peaks, i.e., those reached in the following syllable. Of 322 pre-nuclear pitch accents, 119 were rises (L+H* & L+>H*). These were subset and coded according to *syllable openness* (*open* versus *closed*), and *following nasality* (*nasal* versus *other*). Here, the level *nasal* encompasses any post-vocalic nasal consonant (/n, m, ŋ, ɲ/ in the datset). For the purposes of this paper, the symbol '/N/' is used as shorthand. With regards *syllable openness*, no a priori assumptions were made as to the nature of resyllabification in this dialect, thus syllable structure was coded as citation form, e.g., /βlaɾ/ in *hablar* ('to speak') would be coded as *closed*. The possibility of resyllabification is captured through *following context*, i.e., whether a closed syllable appeared pre-vocalically, pre-consonantly, or pre-pausally. The role of resyllabification

on pitch alignment is not discussed within this paper, but is an avenue for future research.

Nuclear accents were labelled according to the combination of the final pitch accent and boundary tone. Labels were then coded according to whether they were *circumflex* or *other* under the variable *nuclear accent*. The circumflex accent was defined as any accent which completed the peak-and-trough like contour. *circumflex* was therefore considered the following accents: L+ HL%, L+H* L%, L+H* HL%, L*+H HL% (Mota et al. 2011). Any combination outwith this was considered *other*. Nuclear accents were also coded for *length*, i.e., the number of p-words in the utterance: < *1* (1 p-word) and > *1* (2 p-words or more).

Plots were created using *ggplot2* (Wickham 2016) in R (R Core Team 2022). Logistics regression models were run using the *lme4* package (Bates et al. 2015).

4. Results

4.1. Pre-nuclear pitch accent

With all pre-nuclear pitch accents pooled together (N=322), L* and L+H* accounted for the majority of realisations (28.98% each). This is followed by H* and L+>H* which account for 19.97% and 13.07% respectively. H+L* & L*+H occurred least often accounting for 8.13% and 1.03% respectively. Of the 119 rises, L+H* accounted for 68.9% (N=82) and L+>H* 31.1% (N=37).

Specific to peaks, comparisons between open and closed syllables reveal the effect of syllable structure. Tonic peak alignment was more common in closed syllables, where L+H* represented 94.35% of all rises. In open syllables, peaks were more evenly distributed: L+H* accounted for 53.95% and L+>H* 46.05% (see Figure 4). A simple logistics model was run with *ToBI_label* and *syllable openness* as the dependent and independent variable respectively. Results revealed that, relative to L+>H*, L+H* was more likely in closed than open syllables (t = 2.8622, p<.001). Post-hoc pairwise comparisons confirmed these initial conclusions (open-closed, t=-2.86, p<.05).

Data was then subset according to *syllable openness* and *nasality*. Results show divergent peak alignment patterns according to syllable type: in closed syllables with coda /N/, e.g., *descendiente* ('descendent'), 90.05% of peaks were L+H*. In open syllables with /N/ as the following onset, e.g., *mexicano* ('mexican'), 0% of peaks were L+H*; instead, 100% were L+>H*. When the sound was anything other than /N/, the previous effects of syllable aperture emerged: in closed syllables, 100% of peaks were L+H*, whilst in open syllables, this dropped to 69.49%. This is visualised in Figure 5.

It was not possible to statistically test the interaction between *syllable openness* and *nasality* due to the small sample size. Nontheless, the following interaction is noted: in closed syllables with coda /N/, peaks align tonically (L+H*), however in open syllables with the following /N/ onset, post-tonic peak alignment is noted (L+>H*). This therefore suggests that the presence of the post-vocalic nasal plays a unique role in peak offset in this variety.

4.2. Nuclear pitch accent

Of the 120 nuclear accents, *circumflex* configurations accounted for 72% of the dataset (N=86) and *other* 28% (N=34) (see Figure 6). Subset for utterance length, the circumflex accent accounts for the majority of realisations in both utterances of 1 p-word and those of two or more, albeit to a slightly larger extent

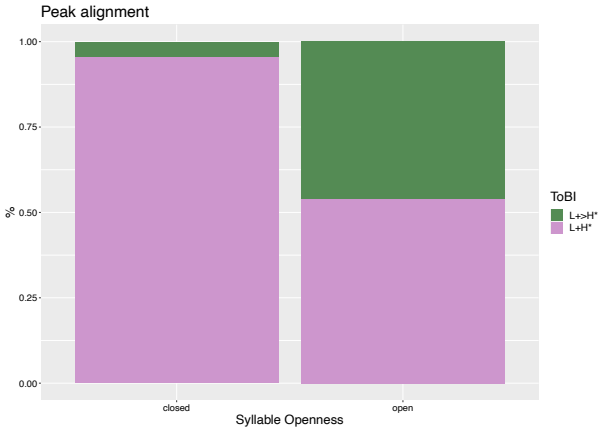


Figure 4: Pre-nuclear peak realisation across closed and open syllables.

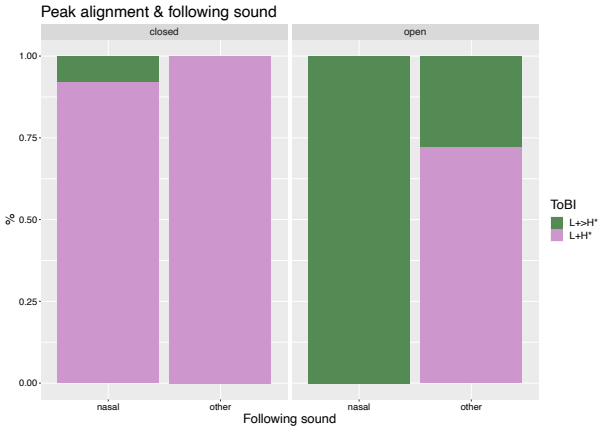


Figure 5: Pre-nuclear peak realisation across nasal contexts and syllable aperture.

in the latter (64.5% and 72.81% respectively) (see Figure 7). A linear mixed effects regression model was run with *nuclear accent* and *length* as the dependent and independent variables respectively. No significant effect emerged (t = -0.685, p > .05), with the circumflex accent as likely in utterances of 1 p-word as in those of 2 or more.

5. Discussion and conclusion

Results present patterns which both align and diverge from those described in non-Afro Mexican varieties. Firstly, L* and L+H* account for the majority of pre-nuclear pitch accents and the circumflex accent in nuclear position. Patterns thus mirror that of phrase-level pitch in non-Afro Mexican Spanishes, where such characteristics signal broad focus, declaratives. Given that pre-nuclear pitch accents are not invariantly high as described in other Afro-Hispanic varieties, this is further indicative of phase-, not word-level, pitch.

Nonetheless, analysis of peak alignment reveals divergent features, specifically an interaction between syllable structure and the segmental string: if present, peaks align on post-vocalic nasals regardless of intervening syllable boundaries. Thus, in closed syllables with coda /N/, peaks align tonically, yet in open syllables with /N/ as the following onset, peaks are post-tonic.

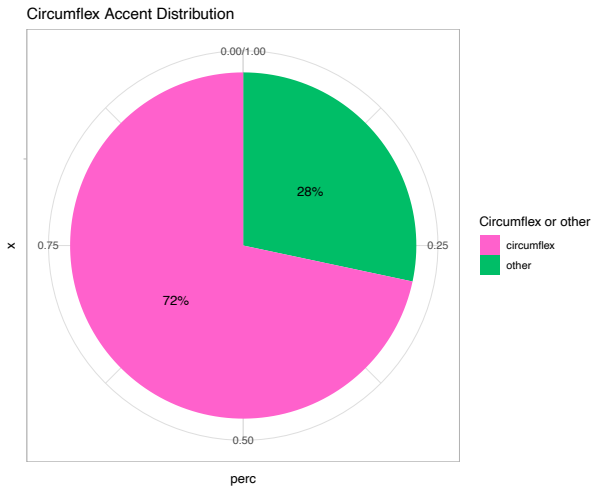


Figure 6: Distribution of circumflex accent versus other in nuclear position.

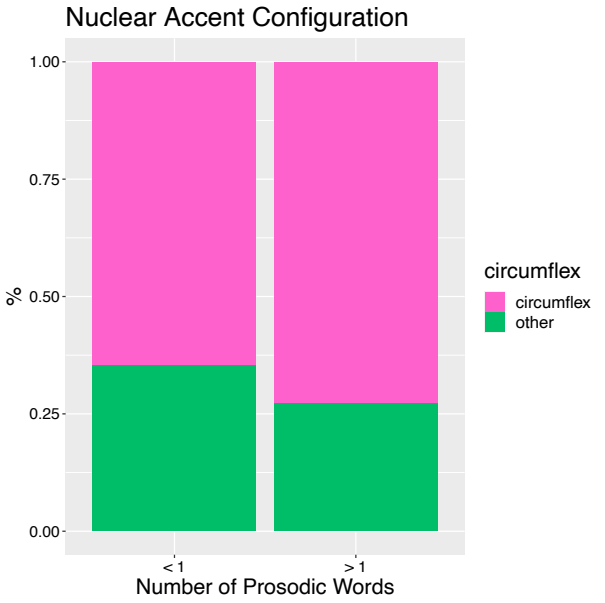


Figure 7: Distribution of circumflex accent versus other in nuclear position according to utterance length.

Outwith the nasal context, tonic peaks are more common in closed syllables than in open. This behaviour is interesting for a number of reasons. Firstly, patterns are distinct from those described both in Afro-Hispanic language and non-Afro Mexican varieties. In the former, similar effects may occur, however due to the the under-researched nature of Afro-Hispanic language, the variability in peak offset is unattested. Moreover, we are unaware of research that has analysed the role of segmental string in non-Afro Mexican Spanishes. Thus it may be that this a common characteristic, however due to the paucity of research in transatlantic varieties, this remains to be explored.

Secondly, these discrepancies raise important theoretical questions surrounding the suitability of the Segmental Anchoring Hypothesis (SAH) in accounting for such behaviour. According to the SAH, tonal movements align with syllabic units, such peak offset should not vary according to syllable structure, nor the segments within. Whilst this may hold for non-Afro Mexican varieties, it is not the case for the data analysed here within. As such, we consider the following options.

Firstly, it may be instead that a lax, dialect-specific SAH emerges due to an underlying phonological feature (Prieto 2009); in this case, either nasality or sonority. For example, whilst both Northern and Southern German speakers showed consistent rise onset patterns, peaks aligned later in Southern German varieties than in Northern (Atterer and Ladd 2004) with similar differences noted between Southern Standard British English (SSBE) and Received Pronunciation English (RP) (Ladd et al. 2009). In this variety, it may therefore be that 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. Again, whether this is specific to Afro-Mexican Spanish requires further comparison with non-Afro Mexican varieties.

Secondly, an articulatory, inter-gestural coordination model may be applicable. It may 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. Such explanations have been used to describe similar differences in tonal offset for both Catalan (Prieto 2009) and Spanish (Prieto, Van Santen, and Hirschberg 1995; Prieto and Torreira 2007). We may therefore theorise that the longer duration of the nasal thus provides a platform for variation in terms of peak alignment, such that instability is noted.

Lastly, a perceptual model may provide an explanation: in order for a rise to understood and interpreted by the interlocutor, it must continue rising *during* the post-vocalic nasal should it occur adjacent to one (House 1990).

These are of course theories at present and require further testing with a larger, more varied dataset. With the goal of assessing the unique role of the nasal and syllabic affiliation, we have run a series of control experiments assessing peak alignment across nasal and sonorous codas and onsets, i.e., i.e., /l, n, s/. Within this, timing measurements will be gathered in order to pinpoint where exactly within the consonants peaks align, and whether this may vary according to the phonetic duration of the consonant itself. Together with the spontaneous speech data, such experiments are also advantageous in order to best control for conflicting influences, e.g., stress clash, adjacency to IP boundaries, and speech rate (Prieto, Van Santen, and Hirschberg 1995; Prieto and Torreira 2007). Comparative data has also been also gathered from speakers of non-Afro Mexican Spanish from Mexico City.

Nonetheless, regardless of the upcoming analyses, the results presented here within are of note. They are indicative that, whilst this variety may not employ pitch at a word-level, unique prosodic features emerge. It therefore highlights the importance of exploring under-researched varieties in order to shine light on theoretical questions surrounding pitch anchoring processes.

6. Acknowledgements

I would like to thank the communities of Cuajinicuilapa and Punta Maldonado for welcoming me and sharing their culture and language with me. I would also like to thank colleagues at the University of Edinburgh and the Universidad Autonóma de Querétaro for their feedback.

7. References

Arends, Jacques and Adrienne Bruyn (1994). "Gradualist and developmental hypotheses". In: *Pidgins and Creoles: An Introduction*. Ed. by Jacques Arends, Pieter Muysken, and Norval Smith. John Benjamins Publishing Company. Chap. 10, pp. 111–120. URL: https://perla.princeton.edu/.

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.

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.

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.

Ladd, D. Robert (2008). "Introduction to intonational phonology'". In: *Intonational Phonology. [Online*, pp. 3–42.

Ladd, D. Robert, Astrid Schepman, Laurence White, Louise May Quarmby, and Rebekah Stackhouse (2009). "Structural and dialectal effects on pitch peak alignment in two varieties of British English". In: *Journal of Phonetics* 37, pp. 145–161.

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

— (2006). "El dialecto afroyungueño de Bolivia: en busca de las raíces del habla afrohispánica". In: *Iberoamericana* 4.2, pp. 137–166.

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

— (2010). *A History of Afro-Hispanic Language Five Centuries, Five Continents*. Cambridge University Press.

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.

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, Chilin Shih, and Holly Nibert (1996). *Pitch downtrend in Spanish*. Tech. rep., pp. 445–473.

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.

Schiel, Florian (1999). "Automatic phonetic transcription of non-prompted speech". In: *ICPhS Proceedings*, pp. 607–610.

Schwegler, Armin (1999). "Monogenesis Revisited: The Spanish Perspective". eng. In: *CREOLE GENESIS, ATTITUDES AND DISCOURSE: STUDIES CELEBRATING CHARLENE J. SATO, Rickford, John R. & Romaine, Suzanne [Eds], Amsterdam: John Benjamins, 1999, pp 235-262*.

— (2001). "The Myth of Decreolization". In: *Degrees in Restructuring in Creole Languages*. Ed. by Ingrid Neumann-Holzschuh and Edgar W. Schneider. John Benjamins Publishing Company, pp. 409–436.

Sessarego, Sandro (2015). *Afro-Peruvian Spanish. Spanish slavery and the legacy of Spanish Creoles*. John Benjamins Publishing Company. URL: http://benjamins.com/catalog/cll.

Torreira, Francisco and Martine Grice (Apr. 2018). "Melodic constructions in Spanish: Metrical structure determines the association properties of intonational tones". In: *Journal of the International Phonetic Association* 48.1, pp. 9–32. DOI: 10.1017/S0025100317000603. URL: https://www.cambridge.org/core/product/identifier/S0025100317000603/type/journal_article.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL: https://ggplot2.tidyverse.org.

Willis, E (2003). "The intonational system of Dominican Spanish: findings and analysis". PhD thesis.

— (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.

Articulatory speech synthesis without phones and gestures?

Konstantin Sering[1], R. Harald Baayen[1]

[1]*University of Tübingen*

konstantin.sering@uni-tuebingen.de, harald.baayen@uni-tuebingen.de

Abstract

With this work we show how speech production can be modelled on the word level without any symbolic units, neither on the acoustic side like phonemes, nor on the semantic side like word types, nor on the motor side like gestures or articulatory targets. We present and discuss a computational model of articulatory speech production, which implements a predictive planning approach, known from hand and arm movements, into the articulatory domain. This computational model is named Predictive Articulatory speech synthesis Utilizing Lexical Embeddings (PAULE). As articulatory speech synthesizer the VocalTractLab speech synthesizer is used, which simulates the human speech system on a geometrical level with 30 different control parameters (channels) and with a time resolution of 401 Hertz. As the synthesis quality of the PAULE shows decent results, we conclude that human speech production can be modelled without the use of any symbolic units like phones and gestures on the word level.

Keywords: speech production, articulatory speech synthesis, predictive planning, motor control, sequence-to-sequence model

1. 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; Sering 2023). 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.

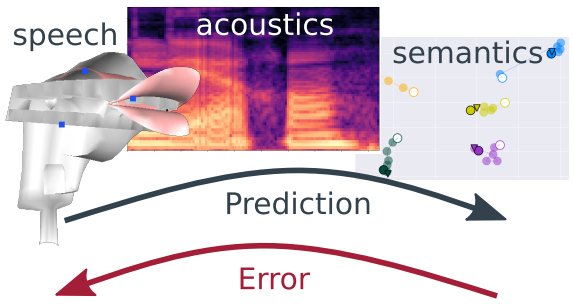


Figure 1: *The predictive principle implemented with PAULE assumes an internal predictive process that predicts the acoustic and semantic effects of an imagined upcoming articulatory motor program.*

[1] https://vocaltractlab.de/index.php?page=vocaltractlab-about

2. Architecture

PAULE[2] implements a predictive planning approach (see Figure 1) 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.

The acoustic and semantic representations are used as goal spaces within PAULE. Planning the cp-trajectories is achieved by minimizing the distance of of the predicted effects to given targets in the goal spaces. The minimization in the goal spaces is done along the local gradients of the forward predictions. Figure 2 depicts this process in a simplified form. Through the exploitation of the local gradients PAULE is capable of optimizing those parts of the cp-trajectory which are perceived as most relevant to the predictive forward model.

PAULE connects the different data structures with learned LSTM-based mappings (Hochreiter and Schmidhuber 1997) (Figure 3). These mappings are pre-trained and back-propagate prediction errors from the semantic and acoustic representations. The 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.

PAULE is implemented and pre-trained to find suitable cp-trajectories for the 4,311 word types of the German language. These can be synthesised by giving the target label semantic vector and a desired duration. Furthermore, PAULE is capable of re-synthesizing longer chunks of of speech signals even from different languages like English in a copy-synthesis setup.

3. 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.

Even the question shifts away from "what is the correct articulation for a given word" to which articulatory patterns are sufficient to satisfy the acoustical and semantic target in mind while complying to some articulatory laziness constraints. The PAULE framework therefore proposes that there is not necessarily a single optimal articulatory control, but a multitude of good controls, which satisfy different goals to different degrees and which is inherently dependent on the perceptive experience of the speaker, her knowledge of the target language and her experience with articulating similar words.

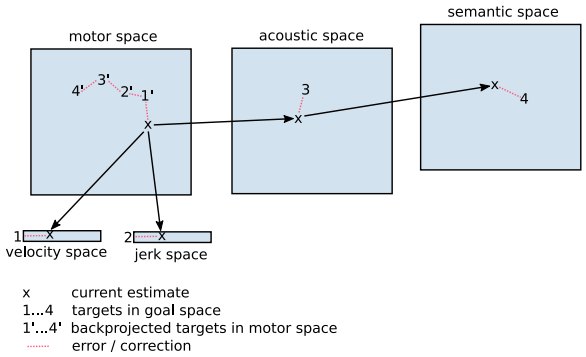


Figure 2: *The predictive principle implemented with PAULE compares the predicted effects in the acoustic and semantic goal space as well as some velocity and jerk constraints. The missmatch (or error) between the predictions and the desired target acoustic respectively semantic representation is used to improve the articulatory motor program along the gradients of the predictions. With this gradient-aware planning (or optimization) only forward models are needed. No explicit model of the error correction is used. Still the error can give locally relevant correctoins. All goal spaces are continuous and therefore no discrete or symbolic representations like phonemes or motor-gestures are used within PAULE.*

4. Discussion and Conclusion

Doing articulatory speech synthesis without any gestures or phones might be seen as a bold claim. But, PAULE is a computational model that does produce control-parameter trajectories for the 30-dimensional articulatory speech synthesiser in VTL on the word-level. PAULE achieves this without the use of any symbolic units in its pipeline.

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

[2] https://github.com/quantling/paule

[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

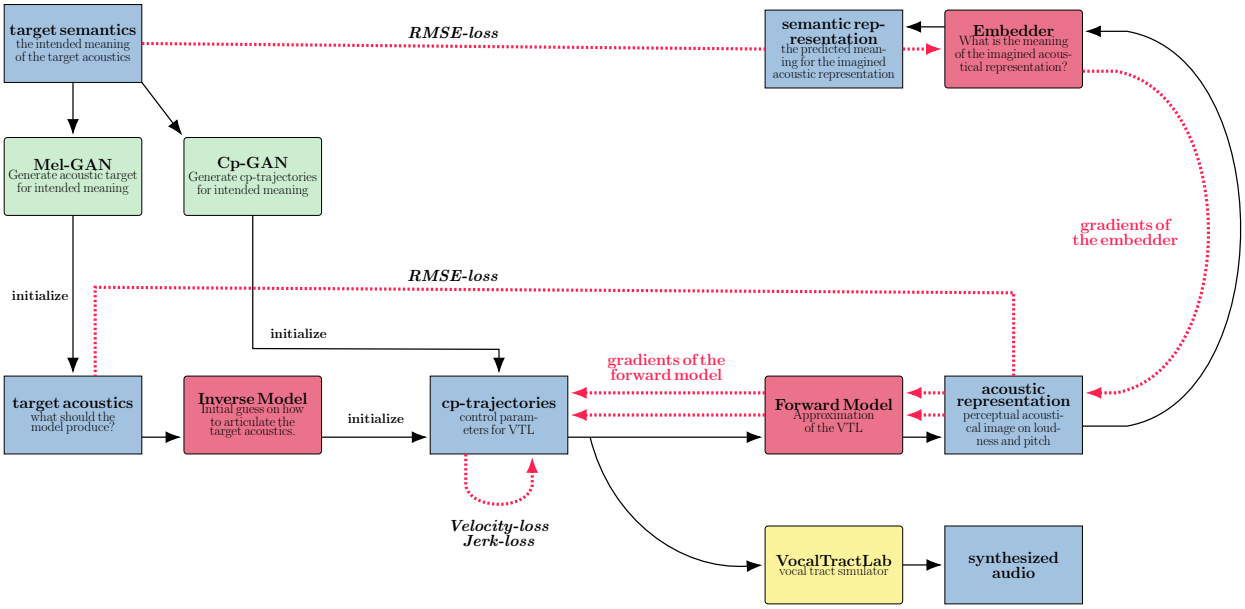


Figure 3: *PAULE plans the articulatory control-parameter(cp-)trajectories of the VocalTractLab geometrical 3-dimensional articulatory synthesizer. The planning uses an internal predictive forward loop and optimizes the upcoming cp-trajectories by minimizing an error in an acoustic and semantic goal space. All data representations uses a fine grained finite time representation as well as a continuous representation for the positions of the articulators as well as for the acoustics features and the semantic lexical embeddings. Therefore, no symbolic representations are needed to synthesize single word tokens with PAULE. The Mel-GAN, Cp-GAN, and the Inverse Model are only used once for initialisation purposes.*

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 real-time 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.

5. Acknowledgements

Konstantin Sering is funded by the DfG grant 527671319 („Komplexe Wörter im Kontext“). Harald Baayen and Konstantin Sering are members of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

6. 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 (2023). *Predictive Articulatory speech synthesis Utilizing Lexical Embeddings (PAULE)*. Tübingen: Universität Tübingen. DOI: 10.15496/publikation-90142.

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.

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, Canada*
[2] *Université du Québec à Montréal, Canada*
[3] *Concordia University, Canada*

eija.aalto@etsmtl.ca, minoru.yoshida.1@ens.etsmtl.ca, menard.lucie@uqam.ca,
walcir.cardoso@concordia.ca, catherine.laporte@etsmtl.ca

Abstract

Ultrasound (US) imaging is a promising visual articulatory biofeedback device for second language (L2) pedagogy, allowing visualization of tongue movements. Despite its potential benefits, uncertainties persist regarding the specific learner profiles that may derive the greatest advantages from US biofeedback. This pilot study aims to provide a means to evaluate L2 learners' ability to effectively utilize visual biofeedback by assessing maximal tongue retraction and lowering immediately before and after a short US biofeedback session. Six participants completed a short learning task, two with previous exposure to US biofeedback and four without. Participants without previous exposure to US biofeedback improved their maximal tongue movements to some extent, while those with previous exposure to US biofeedback showed little improvement. This suggests that this type of task may help characterize learners' receptivity to visual articulatory biofeedback.

Keywords: speech production, ultrasound, biofeedback, L2

1. Introduction

Ultrasound (US) imaging has been used successfully in second language (L2) pronunciation education (Antolík et al., 2019; Bliss et al., 2018; Bryfonski, 2023; Chang, 2023; d'Apolito, 2017). The strength of US as a biofeedback device is its ability to show otherwise invisible internal articulatory movements of the tongue in real time. Unsurprisingly, research reveals that L2 learners welcome this new tool enthusiastically (Bryfonski, 2023; Meadows, 2007; Tsui 2012). For instance, in PICO studies (i.e. participants/population, intervention, comparison, and outcomes), a substantial number of researchers report US being equal to auditory-based methods (Antolik et al., 2019; Chang et al., 2023; Bryfonsk, 2023; Cleland et al., 2015; Lin et al., 2019; Roon et al., 2023). Notwithstanding the comparable outcomes reported in large-scale studies, US biofeedback has demonstrated distinct advantages in certain contexts and for specific learner subgroups. Smaller sample sizes have yielded promising results, with US biofeedback facilitating superior performance compared to traditional methods (d'Appolito et al., 2017; Wu et al., 2015). Moreover, individual differences have been observed, suggesting that some learners may benefit more from the proposed treatment than others (Lin et al., 2019). Notably, US biofeedback has proven advantageous in speech tasks that demand the generalization of learning (Brysonski, 2023), as well as in the discrimination of manner of articulation (Roon et al., 2023). Additionally, US biofeedback has shown particular efficacy in the acquisition of specific target sounds, such as palatal stops (Cleland et al., 2015), and in promoting sustained improvement over time (Incegolu & Gnevsheva, 2020).

Previous literature suggests that learner characteristics may play a pivotal role in influencing learning outcomes in US biofeedback interventions (Chang, 2023; d'Apolito et al., 2017; Li et al., 2019). Individual differences in motor skills, sensory acuity, and other cognitive and physiological factors, are considered potential sources of variability in learning outcomes (Kartushina et al., 2015; Preston et al., 2014). However, studies of L2 learner characteristics remain scarce (Li et al., 2019). Li (2019) reported that while oral sensory acuity did not correlate with learning outcomes when US and auditory methods were employed in L2 interventions, phonological processing skills and variability in pronunciation were found to be significant predictors of learning outcomes, regardless of the teaching method employed. Moreover, Ouni (2014) explored tongue motor control with novel tongue postures as a method of assessing individual oral-motor skills in the presence of US biofeedback. Their results highlighted the efficacy of US as a biofeedback modality for facilitating the acquisition of difficult tongue movement. Furthermore, they noted that speakers often lack awareness of their speech movements, even in their native language.

A critical question that remains unanswered is whether the ability to effectively integrate visual biofeedback of tongue movements is an inherent individual trait, and if so, whether this capacity can be reliably measured. Answering this question would help identify learners who are more likely to benefit from US biofeedback aimed at enhancing their L2 pronunciation skills. The current pilot study investigates one possible way to assess this individual characteristic, based on the speaker's ability to produce maximal tongue retraction and lowering immediately before and after a short US biofeedback session.

2. Methods

2.1. Participants

The participants were six bilingual adult volunteers (2 male, 4 female) without any diagnosed speech, language, communication, cognitive, or memory impairments. Four of the participants (1 male, 3 female) had no previous exposure to US articulatory feedback ("no exposure" - NE group) and two participants had extensive US biofeedback experience ("previous exposure" - PE group).

2.2 Equipment

The US data were recorded with a Telemed MicrUs EXT-1H scanner using an MC4-2R20S-3 transducer operating at a central frequency of 4 MHz. The transducer was kept stable under the participants' mandible by attaching it to a

construction helmet suspender. The participant placed the helmet suspender on their head and secured it to ensure comfortable fitting. The researcher placed the transducer covered with acoustic coupling gel under the participants' jaw and attached elastic bands from the transducer to the helmet suspender, one to the temple and one behind the ear on both sides. The adequacy of the field of view was assessed by ensuring that it covered the tongue surface when producing the syllables [ti] and [ga]. The placement of /t/ was an approximation because the US does not typically show the tongue tip due to the mandible bone and air pocket under the tongue tip. The visibility of the genioglossus tendon was used to confirm that the transducer was placed to correctly produce a midsagittal view of the oral cavity.

2.3. Task

The participant sessions included an introduction to ultrasound imaging and recordings of maximal tongue movements before and after a short ultrasound biofeedback session. All participants received standardized verbal instructions and visual demonstrations of the desired tongue movements, with the researchers modelling the target movements using their hands.

2.3.1. Introduction and pre-biofeedback recording

The participants received a short introduction to US imaging where they were briefly shown how to interpret the image of their mouth cavity. They underwent a practice and recording where they did not see the US screen. They were first asked to orally produce [ti] and [ga]. Then, they performed a warm-up task in which they were asked to keep their teeth clenched to stabilize their jaw and move their tongue in their mouth cavity. Next, maximal tongue movements were requested and recorded. Specifically, the participants were asked to retract their whole tongue as far back as possible and then to lower their whole tongue as low as possible in their mouth while still clenching their teeth.

2.3.2 Short US biofeedback session

After the introduction and initial recording, the US screen was turned towards the participant. They were encouraged to briefly familiarize themselves with the US image by speaking and moving their tongue freely. A short structured tongue movement practice followed where they were asked to practice the maximal retraction and lowering movements while their teeth were clenched. The biofeedback session lasted for a maximum of two minutes including the free speech and guided tongue movement practice.

2.3.3. Post-biofeedback recording

Subsequent to the biofeedback session, a final recording was conducted following the same protocol as the pre-biofeedback recording. The participants were again asked to produce [ti] and [ga] syllables before the tasks of maximal tongue retraction and lowering, this time with their teeth clenched and without the aid of ultrasound biofeedback.

2.4. Analysis

2.4.1 Image registration

Three reference points were extracted manually from the recorded US images. The first of these was the approximate position of the tongue tip during [t], the second was the highest position along the tongue dorsum during [g], and the third was the anterior end of the genioglossus tendon. These reference points were used to approximately co-register the US data from the pre- and post-biofeedback recordings, as illustrated in **Figure 1.** The time points corresponding to maximal tongue retraction and lowering before and after biofeedback were identified by manually searching the synchronized US audio recordings. The corresponding manually traced tongue contours were displayed within the common reference frame obtained through registration.

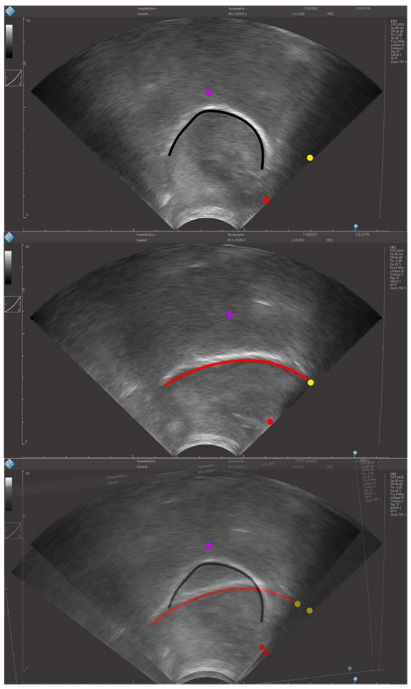


Figure 1: *Co-registration of pre-and post-biofeedback data. 1st image: extracted reference points (yellow- /t/; purple- /g/, red –genioglossus tendon's visible anterior end) and traced tongue contour for pre-biofeedback tongue lowering. 2nd: the post-biofeedback tongue position and the reference points. 3rd: the co-registration of the images. The placement of /g/ was superimposed first and the post-biofeedback image was rotated to align the markings of /t/ and genioglossus tendon.*

2.4.2. Measurement of change in tongue movements

The changes in tongue retraction and lowering were measured from the co-registered tongue contours. Change in tongue retraction was assessed by measuring the distance between the tongue tip's position along the horizontal axis recorded before and after biofeedback, as illustrated in Figure 2. Change in tongue lowering was assessed by measuring the distance between the vertical position of the highest point along the tongue dorsum pre-biofeedback and the corresponding post-biofeedback point of tongue contour on the vertical axis. These measurements are illustrated in Figure 3.

2.4.3. Repeatability analysis

The repeatability of the measurement processes was assessed by comparing the manual registration results for three samples of purposely varying image quality to those obtained by the same operator on the same samples two months later. It was noted that the vertical position of the tongue tip corresponding

to the alveolar ridge [t] and the visible anterior end of the genioglossus tendon were extracted fairly repeatably over the trials. The extraction of the highest tongue dorsum point for [g] was also relatively stable despite slight horizontal movements. Re-extracting the tongue shapes was more prone to error. The main source of error was in the identification of the instant of maximal tongue retraction or lowering. However, the repeatability of distance measurements in the target axis was relatively good despite this.

2.4.4. The comparison of participants' performance

To account for the approximate nature of the co-registration process and facilitate meaningful comparisons, the distance measurements were normalized relative to the data from the participant with maximal observed change and quantized to five bins. The first bin is for the participant that exhibits maximal observed change, whereas participants in the last bin exhibit the least measurable change relative to the maximal observed change. This highlights coarse individual differences in the ability to improve existing tongue movements with a short US biofeedback session.

3. Results

The changes in maximal tongue retraction and lowering from pre to post-biofeedback recordings are illustrated in **Figures 2 and 3.** In the tongue retraction dimension, three of the participants in NE group noticeably increased their tongue retraction after the biofeedback session, while one (NE3) had little change. Both participants in PE group showed little to no change. Four of the participants used the same tongue retraction pattern in pre- and post-biofeedback recordings, while two (NE3 and PE5) produced considerable changes in their tongue shape with minor increases in tongue retraction.

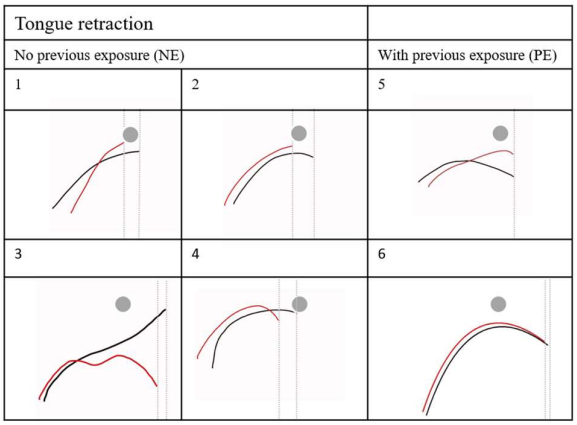


Figure 2: *Maximal tongue retraction before (black line) and after (red line) the US biofeedback practice. The gray dot shows the place of pronounced /g/ in [ga]. The gray vertical lines show the tongue tip position pre and post-biofeedback.*

In the tongue-lowering task, one of the participants (NE2) showed very clear change, one showed moderate change (NE1) and the others showed little change. Two participants changed their tongue shape pattern when attempting to lower their tongue further by curving their tongue tip up in the post-biofeedback recording (NE3 and PE6). In the case of NE3, the tongue tip was actually higher up in the post-biofeedback recording, while the rest of the tongue was slightly lower than in the pre-recording.

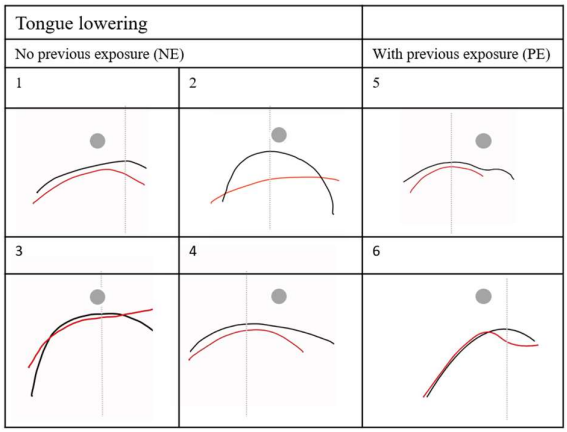


Figure 3: *Maximal tongue lowering before (black line) and after (red line) the US biofeedback. The gray dot shows the place of pronounced /g/ in [ga]. The gray vertical line shows the highest level of tongue dorsum in pre-biofeedback.*

Table 1 shows the relative change between the participants' tongue retraction and lowering from pre- to post-recording. Based on this ranking of change in maximal movements, it seems that participants 1, 2, and 4 in NE group exhibited more noticeable changes in their performance from pre- to post-recording, as compared to the other participants.

Table 1: *Relative change in maximal tongue positions from pre- to post-biofeedback recording. The greatest change was assigned a score of 1, while the other scores were proportionally scaled relative to the largest observed change.*

	NE				PE	
	1	2	3	4	5	6
retraction						
lowering						
Amount of change		100	>75	>50	>25	<25

4. Discussion and conclusion

The current pilot study examined whether a short exposure to US biofeedback can influence maximal tongue retraction and lowering in individuals with and without previous exposure to US biofeedback. The current results suggest that participants without previous exposure to US biofeedback can improve their maximal tongue movements, at least to some extent, with a short US biofeedback session. Participants who received the practice but also had previous exposure to biofeedback showed only negligible to no change. The results also revealed differences between participants' performance, especially in the NE group. Thus, a tongue-shape task with a short US biofeedback may be used as a learner characteristic measure of an individual ability to quickly utilize new visual articulatory information to change their oral motor performance.

The current results align with those of Ouni (2014), who reported improvements in tongue shape accuracy after a US biofeedback session, but not without practice. However, it is important to note that the current results derive from a small sample size (n=6) and lack a comparison or control group, which limits the generalizability of the findings: the observed changes cannot be solely attributed to the effects of US

biofeedback over other types of practice. In addition, one's proprioceptive abilities may play a role in the enhancement of the movements when the same movements are practiced repeatedly. Furthermore, the decision to have participants clench their teeth during the biofeedback recordings served the purpose of isolating tongue movements from those of the jaw, enabling a more focused assessment of their ability to execute the target articulations. However, this methodological choice may have introduced unanticipated challenges for the participants, as the act of lowering the tongue while maintaining a clenched jaw position is counterintuitive and requires overriding habitual motor patterns.

The repeatability analysis of the measurement process revealed challenges in intra-assessor performance. Due to this hardship, the changes in maximal tongue movement amplitude were assessed approximately and relatively between participants. However, this coarse analysis was sufficient to compare the pre- and post-biofeedback change between participants. Thus, the task may provide a much-needed learner characteristic of tongue motor control (Katrushina et al., 2015; Preston, 2014).

From a pedagogical perspective, the short US biofeedback session utilized in this study could be considered a warm-up to phoneme practice. Such a warm-up session may serve to increase the participant's tongue-eye coordination, facilitating the integration of visual articulatory information with the proprioceptive feedback from their speech movements. It could also be leveraged as a diagnostic tool to assess an individual's level of awareness and control over their tongue movements. This information could prove instrumental in elucidating individual differences, or learner characteristics, which may influence the effectiveness of US biofeedback interventions, as suggested by previous research (e.g., d'Apolito et al., 2017; Li et al., 2019).

The findings from this study have opened up several promising avenues for future research. Firstly, the participant's ability to quickly modify their tongue movements with US biofeedback will be used to inform participant group stratification in an ongoing experiment investigating US biofeedback for L2 pronunciation training. Another intriguing avenue would be to compare non-articulatory tongue movements, elicited during the warm-up sessions, not only to speech biofeedback data but also to existing models of speech motor control (Parrell et al., 2019). This new feedback modality, visualization of tongue shape and movement, may affect speech feedback and feedforward mechanisms. While some learners may initially experience confusion when presented with unfamiliar visual articulatory information, others may possess an inherent capacity to readily incorporate the new data into their learning.

5. Acknowledgements

The authors would like to thank all participants. This study was funded by NSERC.

6. References

Antolík, T. K., Pillot-Loiseau, C., & Kamiyama, T. (2019). The effectiveness of real-time ultrasound visual feedback on tongue movements in L2 pronunciation training: Japanese learners' progress on the French vowel contrast /y/-/u/. *Journal of Second Language Pronunciation*, 5(1), 72–97.

Bliss, H., Abel, J., & Gick, B. (2018). Computer-Assisted Visual Articulation Feedback in L2 Pronunciation Instruction: A Review. *Journal of Second Language Pronunciation,* 4(1), 129.

Bryfonski, L. (2023). Is seeing believing? The role of ultrasound tongue imaging and oral corrective feedback in L2 pronunciation development. *Journal of Second Language Pronunciation,* 9(1), 103–129.

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., Nakai, S., & Wrench, A. A. (2015). Helping children learn non-native articulations: The implications for ultrasound-based clinical intervention. *The 18th International Congress of Phonetic Sciences* (ICPhS).

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. *12th International Seminar on Speech Production,* 60–63.

Incegolu & Gnevsheva, K. (2020). Ultrasound Imaging in the Foreign Language Classroom: Outcomes, Challenges, and Students Perceptions. *Pronunciation in Second Language Learning and Teaching Proceedings*, 11(1).

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The journal of the acoustical society of America,* 138(2), 817–832.

Li, J., Ayala, S., Harel, D., Shiller, D., & 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.

Lin, S., Cychosz, M., Shen, A., & Cibelli, E. (2019). The effects of phonetic training and visual feedback on novel contrast production. *The 19th International Congress of Phonetic Sciences* (ICPhS), Melbourne, Australia, 899–903.

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

Ouni, S. (2014). Tongue control and its implication in pronunciation training. *Computer Assisted Language Learning*, 27(5), 439-453

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.

Preston, J. L., McCabe, P., Rivera-Campos, A., Whittle, J. L., Landry, E., & Maas, E. (2014). Ultrasound visual feedback treatment and practice variability for residual speech sound Errors. *Journal of Speech, Language, and Hearing Research,* 57(6), 2102–2115.

Roon, K., Kang, J., Phonetica, D. W.-, & 2020, undefined. (2020). Effects of ultrasound familiarization on production and perception of nonnative contrasts. *Phonetica*, 77(5), 350–393.

Tsui, H. M.-L. (2012). *Ultrasound speech training for Japanese adults learning English as a second language* [Doctoral dissertation, University of British Columbia].

Wu, Y., Gendrot, C., Hallé, P., & Adda-Decker, M. (2015, August). On Improving the Pronunciation of French/r/in Chinese Learners by Using Real-Time Ultrasound Visualization. *The 18th International Congress of Phonetic Sciences* (ICPhS).

Auditory feedback perturbation of F2 in French-speaking children

Isabelle Démosthènes[1,2], Lucie Ménard[1,2]

[1] *Université du Québec à Montréal, Canada*
[2] *Centre de recherche sur le cerveau, le langage et la musique, Canada*
demosthenes.isabelle@courrier.uqam.ca, menard.lucie@uqam.ca

Abstract

In an auditory feedback perturbation, five children and five adults produced the French vowel /ø/ for which F2 in the returning signal was lowered toward /o/. They were also presented with an auditory identification task. Our goal was to investigate if children would show the same pattern of response as adults following a lowering of F2 as they did for different perturbations in previous studies, and if their performance at an identification task, or their production variability would allow us to predict their production in reaction to the perturbation. Surprisingly, the response pattern exhibited by our children didn't follow the recognizable compensatory response displayed by the adults. A bigger sample is needed to determine if this difference can be attributed to variability, to the vowel under study or to the direction of the perturbation.

Keywords*: auditory feedback perturbation, speech production, formant shift, French-speaking children*

1. 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, the exact factors shaping the children's use of sensory feedback and feedforward models are not well understood. In a meta-review of perturbation studies on pediatric populations, Coughler et al. (2022) report on 14 studies that involved real-time perturbation of one or two formant frequencies. In general, results point to the fact that children can produce compensation responses as adults do, but they display larger token-to-token variability. Furthermore, preschool children's responses are also characterized by larger between-speaker variability compared to adults' responses.

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.

More specifically, we will investigate the following questions:

1. Will children show the same pattern of response as adults following a lowering of F2 like they did for other perturbations?
2. Will their performance at an identification task, or their production variability, allow us to predict their production in reaction to the perturbation?

2. Methods

Participants

13 children (age 51-62 months) and 5 adults (age 20-26 years) with no known neurodevelopmental disorder were recruited in Montréal, Canada. Hearing screening at 1000, 2000 and 4000 Hz was carried out on all participants following the Alberta College of Speech-Language Pathologist and Audiologists (2023) protocol. Three children were excluded for not passing the hearing screening and five others due to equipment malfunction (2) or poor data quality (3), leaving us with 5 children (mean 57,6 months) and 5 adults (mean 24,2 years).

Tasks

Participants were presented with two tasks. First, an auditory identification task, displayed using PsychoPy (v2022.2.4) invited participants to select the vowel they perceived between /o/ and /ø/ by clicking on the corresponding picture ("*eau"* /o/, water or "*eux"* /ø/, them). The stimuli were 10 synthesized vowels equally stepped in F2 between the two endpoint stimuli /o/ and /ø/ using the Maeda model (Maeda, 1979). Each stimulus was presented seven times, and all stimuli were presented in random order. Then, in a real-time auditory perturbation task using Audapter (Cai et al., 2008), productions of the vowel /ø/ were gradually 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 heard only their production through the system, white noise was presented in the headphones throughout the perturbation task. To avoid a Lombard effect or discomfort for the participants, a good signal-to-noise ratio has been ensured with the microphone's gain.

Analysis

Identification task data has been analyzed in Matlab (R2022b, Update 7) using the Probit regression method to extract the

slope of the labelling function and the 50% crossover category boundary. For the feedback perturbation task, mean F1, F2 and F3 values have been extracted in Praat (v. 6.1.16) using linear predictive coding in the time interval 20 ms before and after midpoint for each vowel. To allow for intersubject comparison, the frequency obtained for each trial has then been normalized using the formula presented in (1).

$$\frac{trial's\ mean\ formant\ value\ (Hz)}{subject's\ mean\ formant\ value\ during\ baseline\ (Hz)} \quad (1)$$

Ratios around 1 indicate no change 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 used a linear mixed effects model (LMEM) in Jamovi (2.3.28.0) 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 F2 ratio. Finally, for each participant, a perceptual rating of the productions during baseline and hold phases was completed by two blind assessors using a multiple forced-choice interface in Praat.

3. Results

Auditory identification task

Labelling function slope and 50% crossover boundary for the identification task, presented in table 1, revealed significant differences in the perceptual abilities of our two groups. Indeed, linear regression models showed a significant effect of group on slope values ($F(1,8) = 96.7$; $p < 0.001$), but no differences in category boundary values (R^2 0.914). With a mean slope of -1,67 (sd 0.14) and a mean boundary of 5.66 (sd 0.48) adults systematically showed a clearly defined category between /o/ and /ø/. Children, with a mean slope of -0.34 (sd 0.27), showed two different patterns. Two children had no clear distinction between the two sounds, and three exhibited a similar pattern to that of adults, although the boundary was not as clearly defined. The higher slope value for these children might be partially explained by the impulsivity exhibited by some children who quickly gave a response, but then indicated they had wanted to point to the other image. Our results are in line with those of Trudeau-Fisette et al. (in review) who also found a more categorical perception with a steeper slope for adults than children, and no significant difference between groups in terms of category boundary.

Table 1: *Labelling function slope and 50% crossover boundary.*

Group	Participant	Slope	50% boundary
A	1	-1.6	6.16
A	2	-1.63	6.05
A	3	-1.78	5.62
A	4	-1.84	5.49
A	5	-1.49	4.97
C	103	-0.25	8.3
C	104	-0.72	4.8
C	106	-0.39	5.05
C	108	0.02	-0.46
C	109	-0.36	4.49

Perturbation task

Normalized frequencies during experimental trials are presented in Figure 1. The LMEM on F2 ratio showed a significant effect of group ($F(1, 8) = 10.24379$, $p < 0.05$), and a significant effect of the interaction between group and trial ($F(1, 56) = 4.15957$, $p < 0.05$). Surprisingly, no significant effect of phase, either as a main effect or in interaction with group or trial, was found. 33,4% of the variance is explained by this model. This between group difference is clearly visible on figure 1, where on average, adults produce lower F2 ratios than children.

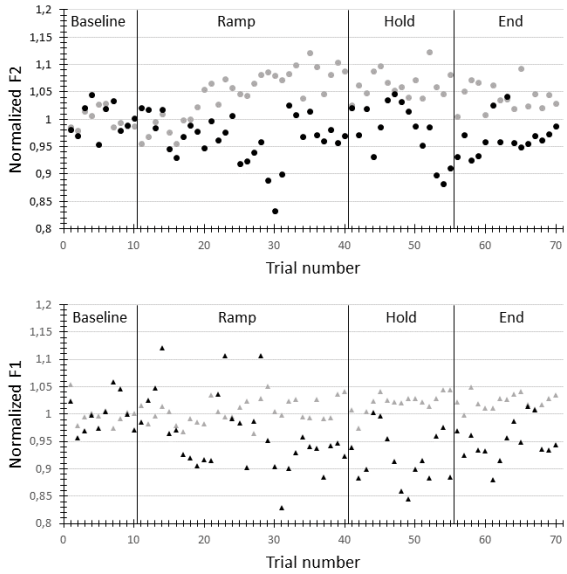


Figure 1: *Mean normalized formant values by phase for /ø/ for adults (grey) and children (black).*
Triangles on the bottom refer to F1 and dots on top to F2.

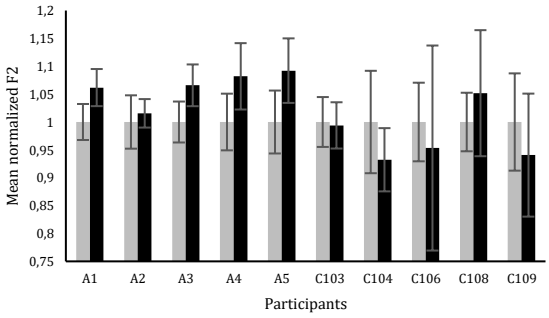


Figure 2: *Mean normalized F2 values for /ø/ during baseline (grey) and hold (black) for each participant.*
Participant numbers beginning with "A" belong to the adult group and those beginning with "C" to the children's group.

A closer inspection of individual data, provided in Figure 2, suggests different patterns of responses between groups. More specifically, when comparing each participant's mean ratio for baseline and hold phase (figure 2), it can be seen that in the adult's group four out of five subjects clearly compensated for the perturbation, where the fifth didn't modify his production much. In comparison, in the children's group, three out of five participants followed the perturbation, one didn't change his production and one clearly compensated. When looking into the effects of the interaction between group and trial on normalized F2, the same trend can be observed: in general, last three trials of a phase are higher for adults, indicating a compensatory pattern, and lower for children, indicating a follower pattern. Although acoustical analysis identified three children as

followers, in the perceptual analysis, only one participant (C106) had productions perceived as /o/ by both raters in the hold phase. This pattern of results will be discussed in the discussion.

These results differ from what had been observed before. Trudeau-Fisette et al. (in review), who applied a perturbation of the same amplitude, on the same vowel, but in the opposite direction (resulting in /ø/ sounding line /e/), mostly had compensating responses in the French-speaking children's group (age 4-6, mean 5y2m). Similarly, MacDonald et al. (2012) applied an F1-F2 perturbation on /ɛ/ making "bed" sound like "bad", had a compensation similar to adults in their English-speaking children's group (mean 51 months).

After visual inspection of each participant's ratio progression throughout phase, some individuals seemed to modify their F1 to compensate for the perceived discrepancy. Also, visual inspection of figure 1, comparing children's F1 ratio to adults', raises the question of a different behaviour between group at F1 level. However, LMEM on F1 ratio showed no effect of group, phase or trial. This is in line with previous studies like Klein et al.'s (2019) who found no consistent effect of F2 shift on F1.

Variability

LMEM on participants' variability level for normalized F2 level, as indicated by standard deviation, revealed a significant effect of group ($F(1, 8) = 5.940$, $p < 0.05$), but no effect of phase either as a main effect or as an interaction with group (figure 3). As seen in previous studies Trudeau-Fisette et al. (in review), our children exhibited more variability than adults, and this was the case throughout the experiment.

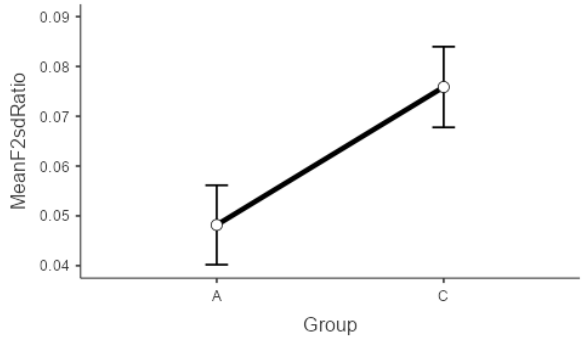


Figure 3: *Mean baseline F2 variability for children and adults as indicated by standard deviation*

The role of perceptual skills and variability

We intended to conduct multiple linear regressions to 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, but the limited number of participants prevents us from doing so. However, when comparing the observed behaviour (follower vs compensator) in figure 2 to the identification slope in table 1, we observe that the three followers (C104, C106 and C109) were the children who had a similar identification pattern to that of adults whereas the only clearly compensating child doesn't exhibit a categorical perception with a nil slope. So, clearly identified phonemic category isn't associated with a compensatory response to perturbation in our children participants. This is consistent with the results of Trudeau-Fisette et al. (in review) who found that, for their children's group, the slope didn't have a significant effect on the level of compensation observed.

Similarly, when we look at the baseline ratio's variability as expressed by the standard deviation (figure 4), we observe that the child who did show a compensation response (C108) and the one who didn't change his production (C103) were those with the lowest F2 ratio variability at baseline. Once more, this is consistent with Trudeau-Fisette et al. (in review) who found that, for children, only F2 ratio variability in the baseline phase had a significant effect on F2 observed in the hold phase.

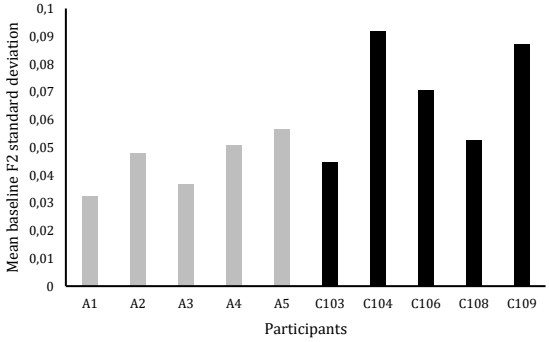


Figure 4: *Mean baseline normalized F2 standard deviation on /ø/ for each participant.*
Participants numbers starting with an "A" belong to the adult group and those with a "C" to the children's.

4. Discussion and conclusion

In this project, our goal was to pursue the investigation of the development of sensorimotor relationship in French-speaking children, by exploring responses to an auditory feedback perturbation in the front-back tongue dimension. To do so, we applied an F2 downward shift to productions of /ø/ in children and adults.

Despite the small sample size, 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. However, contrary to our hypothesis and to what had been previously found in the literature, we didn't find a similar response in both groups: while our adults demonstrated the classical adaptation pattern, our children didn't. To what could have been owed this difference? Three main reasons come to our mind: our limited number of participants, the chosen vowel, and the direction of the perturbation.

This being an exploratory project, with very few participants in each group, it increases the weight of each participant, and one participant with a very different response can change the results. Moreover, considering the high level of variability in children's productions, it is harder to find a clear pattern with a handful of participants.

That being said, our rejection rate in children can be questioned. 61.5% of our participants' results in this group had to be excluded. Comparatively, for a similar age group, it was 16,1%

for MacDonald et al. (2012) and 25.6% for Trudeau-Fisette et al. (in review). How can we explain such a rate difference?

Due to their high pitch, children's formants can be hard to identify and to track. Even MacDonald et al. (2012) who worked on the vowel /ɛ/ which is known to be easy to track had to reject 12,9% of their participants for tracking issues. Hence some rejection at this level is to be expected, and we can expect a higher rate for a vowel that is harder to track like /ø/. Also, during a long and repetitive task like the one we administered, children get bored and find ways to entertain themselves. Many of our participants played with their pitch or their vocal intensity during the ramp or the hold phase. This seems to have affected F2 detection by Audapter and its ability to quickly lower the formant.

Still, Trudeau-Fisette et al. (in review) who worked with the same vowel, /ø/, and a similar age range, did show a lower rejection rate. We think the direction of the perturbation might explain this difference. In their experiment, they increased F2, which led to a drop in intensity of the output signal. The perturbation angle we applied lowered F2, but it consequently increased the intensity of the output signal. Hence, our experiment was more sensitive to vocal intensity increases and generated more signal saturation of the output. Some participants who showed a nice perturbation prior to saturation lowered their voice to a point where the signal wasn't strong enough for Audapter to create the shift afterwards.

Vowel and direction of perturbation come to play not only on the perturbation itself, but also on the response. They are known to be one of many factors affecting responses to perturbation. 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. If the response in this type of perturbation is smaller in adults, could it be later appearing in development?

Similarly, given the articulatory correlates of the different perturbations applied, could articulatory skills be at play here? MacDonald et al. (2012) applied a perturbation that could be, at least partially, compensated by jaw movements, and Trudeau-Fisette et al. (in review) one that could be compensated by lip movements whereas our experiment mainly involved the tongue. If we consider the development of speech motor control and synergies in speech (Namasivayam et al., 2020) the skills needed to counteract the perturbation we applied here, are mastered later than those presented by these authors and hence, associated response patterns could mature later in development.

Finally, although acoustical data provide an objective measure, considering the variability in speech, more data regarding perceptual ratings should be gathered. Indeed, our preliminary results show that only one child produced vowels that were perceived as /o/ despite the fact several children, contrary to adults, produced following responses. This pattern suggests that the same shift might affect speech representations differently across participants. Indeed, children's baseline productions might be in the centre of their perceptual category, such that the feedback shift is not large enough to push the vowel outside the /ø/ category. On the contrary, adults' productions might be located at the periphery of the perceptual target (or their perceptual target might be smaller than that of the children), such that the shift pushes the production outside the /ø/ category, forcing them to compensate for the shift.

Overall, considering our results, we think it is worth pursuing this investigation with a bigger sample of participants and including perturbations along the F2 axis on other vowels.

5. Acknowledgements

We would like to thank two anonymous reviewers for their helpful suggestions.

6. References

Alberta College of Speech-Language Pathologists and Audiologists. (2023). Hearing *Screening Guidelines and Protocol - Preschool to Adult*.

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.

Coughler, C., Quinn de Launay, K. L., Purcell, D. W., Oram Cardy, J., & Beal, D. S. (2022). Pediatric Responses to Fundamental and Formant Frequency Altered Auditory Feedback: A Scoping Review. *Frontiers in Human Neuroscience*, *16*. https://doi.org/10.3389/fnhum.2022.858863

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

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

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

Namasivayam, A. K., Coleman, D., O'Dwyer, A., & van Lieshout, P. (2020). Speech Sound Disorders in Children: An Articulatory Phonology Perspective. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02998

Trudeau-Fisette, P., Vidou, C., & Ménard, L. (in review). *The development of sensorimotor relationships in speech: Adaptation to real-time auditory feedback perturbations*.

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 M. Peterson & Caroline M 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

Abstract

This preliminary study tested the effect of speech therapy on hypophonia and its potential to counteract micrographia in people diagnosed with Parkinson's disease. Data was collected from 3 subjects undergoing speech diagnostics at University of Toledo Speech Clinic. Subjects were asked to speak and write a series of syllables in both a single and dual task paradigm. In the dual task they were asked to speak in their normal, soft, and loud voice. Data was recorded before and after an intensive 12 session speech treatment protocol, where they were trained to speak with intent. Average speech intensity and handwriting stroke area were calculated. Average speech amplitude increased from normal dual task to loud dual task before and after therapy. Following therapy, speech amplitude for soft voice decreased indicating therapy was helpful in modulating amplitude. Handwriting did not show facilitation from speech therapy regarding micrographia. However, variability between repetitions reduced after therapy, showing some coordination between speech and hand movements but coordination is affected by complexity of task and primacy of task.

Keywords: *Parkinson's disease, speech production, handwriting, speech therapy, micrographia*

1. Introduction

The incidence of Parkinson's disease (PD) is increasing rapidly worldwide 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). Hypophonia or reduced speech loudness is a common indication of speech involvement in individuals with PD (Dykstra, 2012). 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 perceive 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 speaking with intent and loudness to address bradykinesia.

There is a gap in the literature regarding the effect of voice on the coordinated movement of hand and speech in PD and the effect of speech therapy on these dual tasks. Therefore, this was a preliminary study designed to investigate the relationship between the dual tasks of speaking and writing, before and after speech therapy, focusing on changes occurring when subjects were asked to speak with a soft or loud voice while performing handwriting.

2. Methods

Handwriting and speech samples were collected from five subjects who participated in the Parkinson's Speech Clinic summer of 2023. One subject was eliminated due to further neurological diagnosis nonindicative of PD, and another due to high cognitive decline that made it difficult to follow the directions of the task, resulting in a sample size of three right-handed males. Handwriting samples were collected before and after 12 intensive speech therapy intervention sessions. All subjects were on their prescribed medication at the time of data collection. Subjects received the SPEAK OUT! therapy protocol where they were trained to speak with "intent". To test the effect that speech has on handwriting, subjects were instructed to write a series of 2-letter syllables on a letter sized unlined white paper using a standardized pen.

Table 1: *Subject Demographics.*

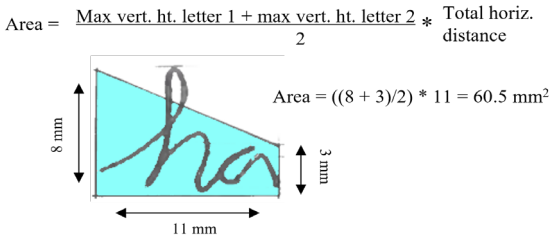
Subject	Gender	Age	Years with PD	Handedness
Subject 1	M	75.6	9.39	Right
Subject 2	M	70.4	4.39	Right
Subject 3	M	79	8.40	Right

Syllables analyzed were "ha", "li", and "te". For the single task procedure, subjects were instructed to write each syllable five times without voicing (single task control for writing) and

similarly to voice without writing (single task control for voice). For a dual task procedure, subjects were instructed to write and enunciate the target syllables in their normal voice (dual task control normal), loud voice (dual task loud), and soft voice (dual task soft). Instructions for loud and soft voice were provided in a randomized block design. Subjects were given breaks as requested.

To analyze handwriting, writing samples were magnified 400x to enhanced measurement precision. The largest stroke of each 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. Due to extreme variation from individual subjects and between writing samples before and after therapy by the same subjects, 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 (Fig. 1). The calculated area was then used for the analyses instead of individual stroke lengths. A 2-way repeated measures analysis was performed for handwriting area using a mixed model to account for missing values. Where main effects were detected comparing before and after therapy, a Tukey test was used to correct for multiple comparisons and detect where those differences were within tasks. Significance was set at $P \leq 0.05$.

Figure 1: *Handwriting example with measurements and area calculation.*



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)). Average intensity measurements were made for each target syllable and averaged over the repetitions. Averaged intensity values were analyzed in a univariate analysis. The post-hoc Tukey test was further conducted to distinguish significant differences between tasks. Significance was determined at the level of $P \leq 0.05$.

3. Results

The average speech intensity was calculated for all syllables separated by voice and therapy conditions. Average intensity varied from speaker to speaker depending on the number of years of diagnosis (Table 1 and Fig. 2). Subjects 1 and 3 (greater than 5 years post diagnosis) performed significantly ($F_{(1)} = 7.9$, p=.006; $F_{(1)} = 8.7$, p=.004) worse after speech therapy than before speech therapy. While subject 2 with the least number of years of diagnosis performed significantly better following speech therapy ($F_{(1)} = 8.15$, p= .005). However, for all subjects, speech intensity significantly decreased from the single task condition to the dual task condition for normal voice (Table 2). PD subjects were able to differentiate between soft voice and loud voice and this distinction was further facilitated by speech therapy (Fig. 2). Univariate analysis revealed significant differences before and after speech therapy, and between all tasks. However, there was no significant interaction between task and treatment condition.

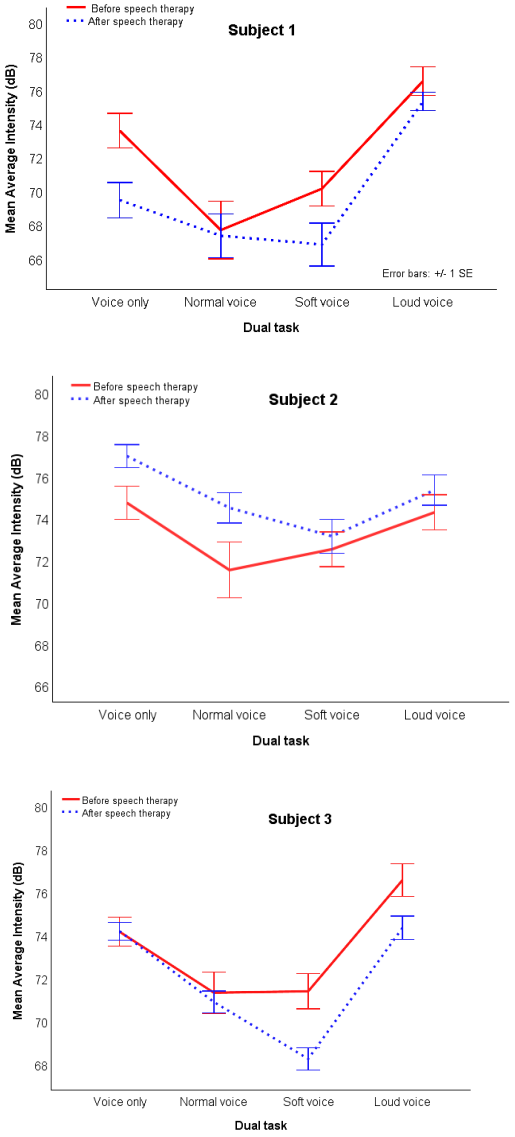


Figure 1: *Line graphs depicting changes in voice intensity for the conditions of single task voice only, dual task normal voice, dual task soft voice and dual task loud voice before and after speech therapy for each subject.*

Table 2: *Mean intensity difference between single task and dual task (P values) when compared to single task for all subjects. Colored boxes indicate significance at $P \leq 0.05$.*

Subject	Dual task Normal voice	Dual task soft voice	Dual task loud voice
Subject 1	4.0 (0.003)	2.9 (0.05)	-4.4 (<0.001)
Subject 2	2.8 (0.006)	3.0 (0.004)	1.0 (0.614)
Subject 3	3.2 (<0.001)	4.3 (<0.001)	-1.3 (0.17)

Average trapezoidal area for the written syllables were calculated separately for voice condition, before and after speech therapy. Handwriting area was reduced following therapy for all subjects (Fig 3). This main effect was

significant ($p \leq 0.05$) for all 3 subjects. For subject 1 a significant reduction was then detected in dual task normal and soft voice and in subject 2 a significant

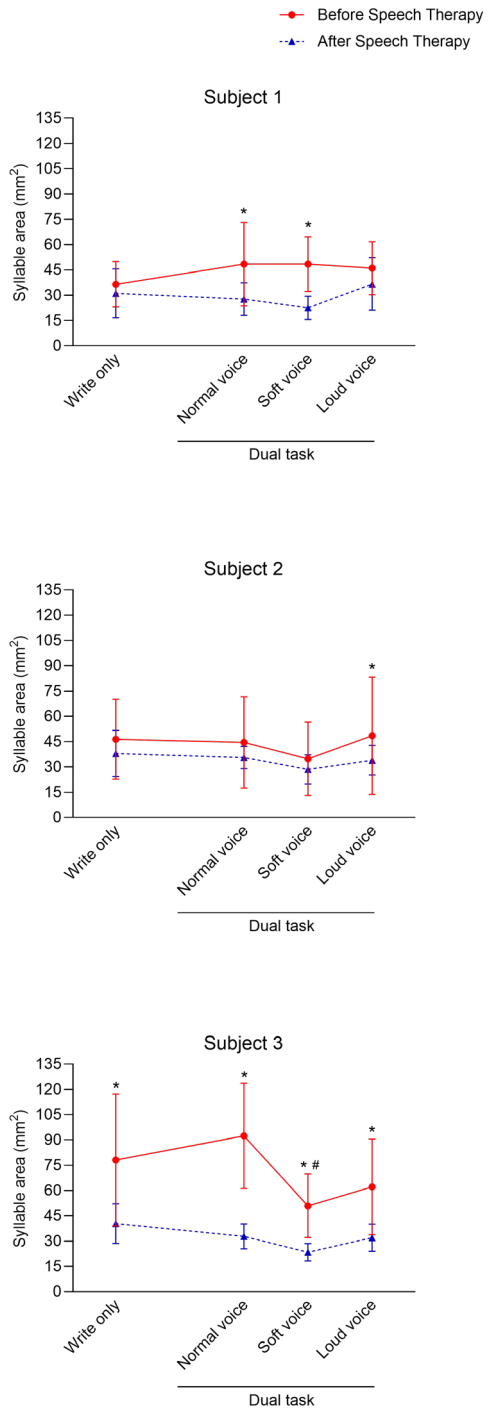


Figure 3: *Line graphs of handwriting area averaged for all syllables before and after speech therapy separated by subject. Data presented as mean ± SD. *Task different before and after speech therapy. #Write only and soft voice different before therapy.*

difference was only detected with dual loud voice. Otherwise, these subjects modulated their writing area similarly before and after therapy. Missing values for normal voice in subject 3 before speech therapy may have accounted for the large difference detected with that condition, however significant reductions in area were detected with all conditions in that subject.

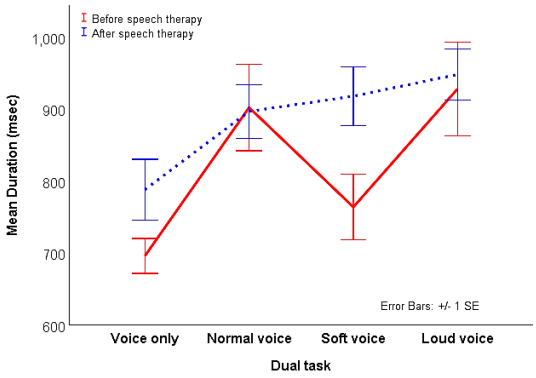


Figure 4: Average acoustic syllable duration combined for all subjects, separated between tasks before and after speech therapy.

Figure 4 displays results of average acoustic syllable combined for all subjects. Durations in the dual task were longer than those of the single task. Furthermore, post-therapy durations values were longer except for dual task normal voice condition. It was significantly longest in the soft voice condition again revealing that therapy was beneficial in distinguishing soft voice from normal and loud voice. In effect, therapy increased the range of duration.

4. Discussion and conclusion

Subjects handwriting and voice were distinctly different for the different conditions measured here. Soft voice condition revealed a loudness level lower than normal, while loud voice condition resulted in increased loudness level regardless of treatment. However, no significant differences were observed in writing area. The results indicate that speech treatment had no faciliatory effect on handwriting.

Speech treatment produced different results on the subject's vocal amplitude, with Subject 2 being the only participant who exhibited increased volume following therapy. However, therapy was helpful in modulating speech intensity such that soft voice had a larger decrease in amplitude when compared to loud voice which revealed an increase in amplitude. While it might be surprising that therapy did not increase voice amplitude for all task, one needs to remember that the SPEAK OUT! treatment protocol addresses the speaker's intentional speech rather than loudness. Therefore, better modulation of volume and increased range of vocal amplitude is the desired outcome of SPEAK OUT!

While vocal amplitude did not change for all speakers following therapy, all speakers increased syllable durations in the post-therapy condition. Most interesting is the duration of the soft voice dual task in the post-therapy condition. Here we see that durations have been significantly increased compared to the pre-therapy condition. This indicates that subjects took longer to say the syllable even when the syllables were produced with lower intensity values.

It is not clear why handwriting area decreased following therapy, but some understanding might be gleaned from the behavior for both voice and handwriting between the dual task of loud voice and the single task (voice or writing alone). In the single task, both voice and handwriting were relatively good, but in the dual task, voice goals appear to be prioritized over handwriting confirming the findings of van Gemmert (1999).

Further, observations of the raw data also showed evidence where the syllable was voiced repeatedly more than the written output. Handwriting and voicing were also not synchronously produced, with voice production often leading handwriting. As voice amplitude increased, speech duration also increased but handwriting area decreased.

Another observation was that variation in writing syllable repetitions was greater before therapy, indicating a potential improvement not detected with our analysis. Instructions were not given regarding use of upper- or lower-case letters and print or cursive writing, resulting in varying handwriting styles between subjects and from writing before and after therapy by the same subject. 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.

5. Acknowledgements

The authors would like to thank all the clients who participated in this study.

6. 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.

Dykstra, A., Adams, S., & Jog, M. (2012). The Effect of Background Noise on the Speech Intensity of Individuals with Hypophonia Associated with Parkinson's disease. *Journal of Medical Speech-Language Pathology*, 20(3), 19–30.

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 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 Behavior. 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.

Perception of a four-way stop laryngeal contrast in Eastern Oromo

Maida Percival[1]

[1]*University of Toronto, Canada*
maida.percival@mail.utoronto.ca

Abstract

The goal of this paper is to identify perceptual cues to the four-way coronal stop contrast in Eastern Oromo. 25 listeners took part in a forced choice identification task with a real-word minimal quadruplet where the closure, burst, and following vowel were systematically alternated and where the burst intensity was manipulated. Results indicated that listener responses were influenced by each acoustic dimension, but that the burst was the most important cue. The other cues were used less consistently, and tended to group based on voicing or constricted glottis status.

Keywords: speech perception, Eastern Oromo, stop laryngeal contrast

1. Introduction

Eastern Oromo 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. This paper examines the perceptual cues to this laryngeal stop contrast, where ejective stops, implosives, voiced pulmonic stops, and voiceless pulmonic stops contrast, as shown in the stop inventory in Table 1. It focuses on the singleton versions, and asks which aspects of the acoustics of each stop laryngeal type are used by listeners as cues to differentiate between the different singleton coronal stops in perception. 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). This has created a puzzle for phonologists in determining how to differentiate two [+constricted glottis, - voice] segments. In phonetics, this is more straightforward, as there are differences in articulation due to the differing airstream mechanism used to produce the ejective and the implosive. Nonetheless, work examining the production of coronal stop contrasts in Oromo has found that the implosive phonetically varies in voicing (Percival, Kochetov, and Kang 2018; Percival 2018). An additional way of investigating how voicing and airstream are maintained in this stop contrast is perception, which is the contribution of this paper.

Table 1: *Eastern Oromo stop inventory (in IPA). Adapted from Owens (1985).*

Stop type	Bilabial		Coronal		Velar		Glottal
voiced	b	b:	d	d:	g	g:	ʔ
voiceless	(p)		t	t:	k	k:	
ejective	p'	p':	t'	t':	k'	k':	
implosive			ɗ	ɗ:			

2. Methods

2.1. Participants

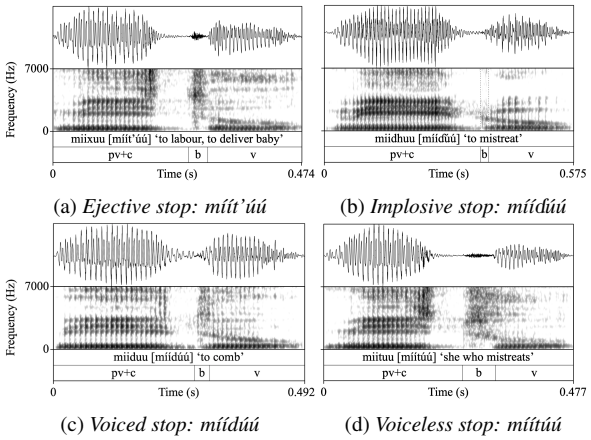
25 first language speakers of Eastern Oromo (12 female and 13 male, aged 17-62 (mean = 42, SD = 16)) participated in the experiment. 21 participants were born in the Oromiya region of Ethiopia and spoke Eastern Oromo as their first language. There were also four participants born in or living in Canada since infancy, who were bilingual in Oromo and English and spoke Oromo as a heritage language. Ethiopia is a multilingual country, and many participants had also lived in other countries prior to immigrating to Canada, where all participants resided at the time of the study. As such, besides English, many participants spoke a number of additional second languages, namely: Amharic (12), Arabic (12), Somali (11), Italian (6), Harari (4), and Swahili (2). No participants reported hearing loss.

2.2. Materials

The stimuli were created from a naturally produced minimal quadruplet which differed only in the word-medial coronal stop present: [míít'úú] 'to labour, to deliver baby', [míítúú] 'she who mistreats', [míídúú] 'to comb', and [míɗúú] 'to mistreat'. Multiple repetitions of these words were recorded in Audacity (Audacity Team 2018) using a Blue Yeti microphone with a sample rate of 48 kHz, and a bit depth of 16 bits. The recording session took place in a quiet room at an Oromo community centre in Toronto, Canada, and the speaker was a 62 year old male first language speaker of Eastern Oromo. One repetition of each recorded word was chosen to be the baselines from which the stimuli were created. The main factor in the choice of repetition was to prioritize the repetitions with the least background noise, and a neutral sounding intonation.

The stimuli were manipulated using Praat (Boersma and Weenink 1992−2024). Each of the four words were divided into three pieces, 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 division points between the pieces were taken to be the first zero crossing before the onset of the burst and the first zero crossing prior to the onset of voicing after any spikes or noise associated with the bursts. Within the burst piece, the mean intensity was manipulated such that there were two versions of each baseline burst: quiet (55 dB) and loud (65 dB). This difference of 55 and 65 dB were chosen based on the values of the first and third quartile of mean burst intensity values for the speaker, rounded to the nearest 5, across all recorded productions of the stimulus list. Another factor in the choice of intensity was that a 10 dB difference perceptually corresponds to twice as loud. After the bursts' intensity was manipulated, the three pieces of each word were systematically alternated to create all combinations. This resulted in a total of 128 stimuli (4 baseline preceding vowel

and closures x 4 baseline bursts x 2 burst intensities x 4 following vowels).



(a) *Ejective stop: míít'úú* (b) *Implosive stop: míídúú*

(c) *Voiced stop: míídúú* (d) *Voiceless stop: míítúú*

Figure 1: *Baseline tokens for the stimuli*

2.3. Procedure

The stimuli were presented to listeners in a forced-choice identification task made using jsPsych (de Leeuw 2015). Participants were instructed to listen to each of the stimuli, and, for each, to click on the word that they heard. They were shown each word choice on the screen, illustrated with a representative image and text in Oromo orthography. They also had the option to replay a stimulus in case they did not hear it the first time, but this option was seldom used. The experiment took place online during the Covid-19 pandemic, and as such participants used their own devices and headphones to participate. In cases where participants did not have access to headphones, they listened without them.

Because the participants in this study were not used to participating in academic research, a number of practices were included to make sure that they would be able to complete the task successfully. First, they were played a sample recording of the minimal quadruplet (a different recording from that used as the baseline) prior to beginning the experiment to make sure that they were familiar with the words. Next, they were given a practice set of natural (non-manipulated) recordings of a different word set to familiarize themselves with the procedure. Finally, in order to prevent possible discomfort or confusion when hearing manipulated speech of their language, the following was included in writing on the experiment instructions page: "There are no right or wrong choices, rather you can think of yourself as a teacher helping your student to learn the difference between similar sounding words. If you're not sure what word you're hearing, don't overthink it and choose the word that sounds closest to what you hear."

2.4. Statistical analysis

Statistical analyses were performed on the 3200 listener responses (128 stimuli x 25 participants) to test the extent of listeners' reliance on each acoustic dimension in the perception of each stop type. The analyses were in the form of generalized linear mixed effects models and were performed in R (R Core Team 2017-2024) using the lme4 package (Bates et al. 2015). There were four models total, one for each response type (ejective, implosive, voiced, voiceless), and they took the form of: response (1 for the response type of interest, 0 for the other response types) $\sim$ burst type $+$ burst intensity $+$ vowel type $+$ preceding syllable & closure type $+$ (1 $|$ Participant). For each model, the response variable was the participants' response type, and the predictor variables were the dimensions of manipulation that the stimuli had undergone. All predictor variables in each language were simple coded (e.g.: low burst intensity = -0.5, high burst intensity = 0.5). Random intercepts were included for Participant. Random slopes were not included because most models failed to converge when they were included, and so rather than have different random effects structures for each model, a simpler random effects structure that could be consistent across all models was included instead. The lack of random slopes means that the results may be anticonservative (Barr et al. 2013) and should be taken with caution. However, when comparing the models to equivalent non-converging models with random slopes, the presence or absence of random slopes did not seem to affect the patterns or significance of the predictors. A p-value of less than 0.05 was taken as significant.

3. Results

The mean percent response for each manipulated dimension is illustrated in Figure 2. In this figure, each manipulated dimension is presented in a separate panel, and the results of the statistical models are presented in Table 2.

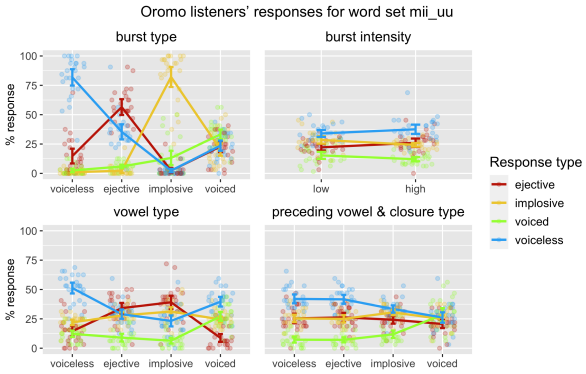


Figure 2: *% response for each response type by dimension of manipulation. % response is on the y-axis, response type is indicated through line colours, and levels of each dimension are given on the x-axis. Error bars represent two standard errors from the mean, and dots represent participant means.*

3.1. Burst type

Across the dimensions, burst type showed the largest effect on listener responses, particularly on the % ejective (red), implosive (yellow), and voiceless pulmonic (green) 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. The results of the statistical models confirm these patterns.

3.2. Burst intensity

As seen in Figure 2, listeners were slightly more likely to respond ejective (red) or voiceless pulmonic (blue) with high burst intensity, and more likely to respond implosive (yellow) or voiced (green) with low burst intensity. This pattern was found to be statistically significant, as seen in the results of the models in Table 2.

3.3. Following vowel type

With vowel type, listeners responded with a given response type significantly more when that baseline vowel type was present. For example, there were more voiced responses when the baseline voiced vowel was present than with any of the other baseline vowel types. The exception to this is that, surprisingly, listeners responded ejective (red line) more when the baseline implosive vowel was present than when the baseline ejective vowel was present. These findings were found to be significant in the models in Table 2.

Table 2: *Statistical results of models for each response type. Significant effects are taken to be at the p<.05 level. Italics = reference level. ej = ejective, vl = voiceless, vd = voiced, im = implosive*

	β	SE	z	p
% ejective response ~ burst type + burst intensity + vowel type + preceding vowel and closure type + (1 | Participant)				
Intercept	-2.02	0.16	-12.39	<0.001
Burst type (*ej* vs. vl)	-2.65	0.15	-17.81	<0.001
Burst type (*ej* vs. vd)	-2.00	0.14	-14.75	<0.001
Burst type (*ej* vs. im)	-4.78	0.26	-18.36	<0.001
Burst intensity (*low* vs. high)	0.35	0.11	3.27	0.001
Vowel type (*ej* vs. vl)	-1.66	0.16	-10.65	<0.001
Vowel type (*ej* vs. vd)	-2.39	0.18	-13.56	<0.001
Vowel type (*ej* vs. im)	0.35	0.13	2.62	0.009
Closure type (*ej* vs. vl)	-0.08	0.15	-0.52	0.607
Closure type (*ej* vs. vd)	-0.51	0.15	-3.38	<0.001
Closure type (*ej* vs. im)	-0.17	0.15	-1.18	0.237
% implosive response ~ burst type + burst intensity + vowel type + preceding vowel and closure type + (1 | Participant)				
Intercept	-2.24	0.17	-12.70	<0.001
Burst type (*im*.vs. vd)	-3.35	0.15	-22.11	<0.001
Burst type (*im* vs. vl)	-6.75	0.38	-17.63	<0.001
Burst type (*im* vs. ej)	-5.91	0.28	-21.35	<0.001
Burst intensity (*low* vs. high)	-0.55	0.13	-4.23	<0.001
Vowel type (*im*.vs. vd)	-0.87	0.18	-4.78	<0.001
Vowel type (*im* vs. vl)	-1.21	0.19	-6.5	<0.001
Vowel type (*im* vs. ej)	-0.43	0.18	-2.39	0.017
Closure type (*im*.vs. vd)	-0.74	0.18	-4.08	<0.001
Closure type (*im* vs. vl)	-0.66	0.18	-3.64	<0.001
Closure type (*im* vs. ej)	-0.71	0.18	-3.91	<0.001
% voiced response ~ burst type + burst intensity + vowel type + preceding vowel and closure type + (1 | Participant)				
Intercept	-2.99	0.17	-17.81	<0.001
Burst Type (*vd* vs. im)	-1.65	0.16	-10.60	<0.001
Burst Type (*vd* vs. ej)	-2.76	0.20	-13.88	<0.001
Burst Type (*vd* vs. vl)	-3.78	0.27	-14.15	<0.001
Burst intensity (*low* vs. high)	-0.42	0.13	-3.29	0.001
Vowel Type (*vd* vs. im)	-2.43	0.20	-12.06	<0.001
Vowel Type (*vd* vs. ej)	-1.94	0.18	-10.67	<0.001
Vowel Type (*vd* vs. vl)	-1.46	0.17	-8.78	<0.001
Closure Type (*vd* vs. im)	-1.60	0.17	-9.6	<0.001
Closure Type (*vd* vs. ej)	-2.35	0.19	-12.17	<0.001
Closure Type (*vd* vs. vl)	-2.33	0.19	-12.11	<0.001
% voiceless response ~ burst type + burst intensity + vowel type + preceding vowel and closure type + (1 | Participant)				
Intercept	-1.19	0.14	-8.54	<0.001
Burst Type (*vl* vs. ej)	-2.68	0.14	-18.95	<0.001
Burst Type (*vl* vs. im)	-6.30	0.28	-22.61	<0.001
Burst Type (*vl* vs. vd)	-3.41	0.15	-22.31	<0.001
Burst intensity (*low* vs. high)	0.30	0.10	2.94	0.003
Vowel Type (*vl* vs. ej)	-1.79	0.15	-11.83	<0.001
Vowel Type (*vl* vs. im)	-2.34	0.16	-14.54	<0.001
Vowel Type (*vl* vs. vd)	-0.90	0.14	-6.35	<0.001
Closure Type (*vl* vs. ej)	-0.03	0.14	-0.21	0.831
Closure Type (*vl* vs. im)	-0.72	0.15	-4.92	<0.001
Closure Type (*vl* vs. vd)	-1.38	0.15	-8.99	<0.001

3.4. Preceding vowel and closure type

The results for preceding vowel and closure type are shown in Figure 2 and presented in the models in Table 2. 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 in Figure 2 this effect appears to be greater with voiceless stops and ejectives than with implosives. Voiceless pulmonic responses (blue) showed an opposing pattern: they decreased significantly with voiced and implosive stop preceding vowels and closures than with ejective and voiceless pulmonic preceding vowels and closures, though Figure 2 seems to show the decrease in % voiceless response to be less with implosives than voiced stops.

4. Discussion and conclusion

Each of the acoustic dimensions manipulated in the stimuli were found to be used in the perception of the four-way stop contrast in Eastern Oromo. Burst type seemed to be a primary cue to the stop laryngeal contrast, particularly for voiceless, implosive, and ejective stops, given the large percent of each response type elicited by the corresponding baseline burst as opposed to other burst types. The results also suggest that the other dimensions may be secondary cues, but whose use is affected by perceptual similarities in the stop types across certain phonetic features.

Table 3 summarizes the acoustics of the baseline tokens for different acoustic intervals, based on qualitative examination in Praat (which can also be seen in Figure 1) and measurements of duration. The importance of the burst type as a cue to the stop contrast is perhaps not unexpected given that the baseline bursts are quite distinct from one another in acoustics. The implosive burst is shorter and less distinct compared to the other burst types as the other bursts seem to have somewhat affricated releases, likely as a result of being followed by a high back vowel. The ejective and voiced bursts are similar in duration, but the voiced burst is produced with periodicity throughout, while the ejective burst is completely voiceless, as a result of the glottal closure which characterizes this stop type. The voiceless pulmonic burst is also voiceless, but the burst produced with some aspiration, reflecting that this stop type is produced with an open glottis (not closed, like ejectives). This results in a longer burst duration than any of the other stop types.

The use of burst intensity and preceding vowel and closure type as cues seem to be influenced by the presence versus absence of phonetic voicing. Listeners associated low intensity bursts with implosive or voiced stops (both with phonetically voiced bursts, as seen in Figure 1 and summarized in Table 3). In contrast, they heard high intensity bursts as either ejective or voiceless pulmonic stops (both with phonetically voiceless bursts). This result is unexpected, as Oromo voiced stops have been found to have higher intensity bursts than voiceless and ejective stops in production (Percival 2014). These effects were small, but they suggest that intensity may be a secondary cue to voicing in Oromo. As for preceding vowel and closure type, listener responses did not differ in perception between stimuli

Table 3: *Stimuli acoustics summarized by acoustic interval*

Interval	míít'úú	míítúú	míídúú	míɗúú
Preceding vowel	0.149 s, spirantized offset	0.145 s, spirantized offset	0.167 s	0.214 s
Closure	0.060 s, no voicing	0.053 s, no voicing	0.045 s, voicing	0.060 s, low amplitude voicing
Burst	0.037 s, slightly affricated	0.063 s, slightly affricated & aspirated	0.030 s, voicing	0.019 s, slightly affricated, voicing
Following vowel	0.159 s, creaky onset	0.134 s, modal onset	0.170 s, modal onset	0.182, creaky throughout

with baseline voiceless closures (whether ejective or voiceless pulmonic), but they heard more voiced stops for stimuli with baseline voiced preceding vowels and closures. In addition, based on Figure 2, baseline implosive stop closures also seem to be more likely to be heard as voiced than ejective or voiceless pulmonic preceding vowels and closures, and the implosive baseline token was produced with partial and/or low intensity voicing. The stops with voiceless closures also seem to have a brief period of frication at the offset of the preceding vowel, which is likely due to the presence of high vowels in the word, but which may have been another cue to voicing within the preceding vowels and closures.

The use of following vowel type as a cue seems to be influenced by the airstream mechanism of the baseline stop, or the presence versus absence of constricted glottis. Listeners broadly responded with either glottalic stop type more when either baseline glottalic stop's vowel type was present, and in contrast responded with either pulmonic stop more for either baseline pulmonic stop vowel type. There was also a small difference in the use of vowel type between ejective and implosive stop types. One unexpected finding was that 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 and summarized in Table 3, the baseline implosive vowel has the most extensive creaky voice of the baseline vowel types. It extends throughout the whole vowel, while the ejective baseline vowel is creaky at the voicing onset but becomes modal by the vowel midpoint. Given that the baseline implosive vowel ended up eliciting more ejective responses, this suggests that creaky voicing may be a slightly more important cue to ejectives than to implosives.

As for the status of the implosive, perceptually implosives were not found to pattern fully as voiced segments do. While they group with voiced stops in the use of burst intensity, they group in-between the two voiceless stop types and the voiced stops in terms of preceding vowel and closure type. Furthermore, they did not act like voiced stops in the use of burst type or vowel type. The implosives and ejectives relying on cues differently or to different extents suggests that even if both segments phonologically pattern as voiceless and constricted glottis, perceptually listeners are still sensitive to voicing differences (as well as potential differences related to airstream).

There are a number of areas of future research for this project. Examining the acoustic cues to the stop contrast across a wider range of words and environments (for example in different positions or different vowel contexts) is one such area, given that the present study consisted of only word-medial tokens where preceding vowels could provide cues, and given that the vowel context was all high vowels, which likely introduced additional cues such as periods of frication. In addition, the present study was limited in that it systematically alternated pieces of stimuli without testing individual acoustic events or measurements within each piece. As a follow-up, investigating the effects of additional single acoustic dimensions within each of burst type, vowel type, and preceding vowel and closure type could clarify which aspects of each of these intervals are the most important to perception. In particular, manipulating different acoustic correlates to creaky voicing within the vowel onset would be interesting given that there are different types of creaky voice which vary in their acoustic correlates (Keating, Garellek, and Kreiman 2015) and it is not clear if listeners associate certain types more than others with ejective or implosive stop contrasts.

5. Acknowledgements

Thank you to Abdulhamid Ahmed, Tamam Youssouf, my dissertation committee (Dr. Jessamyn Schertz, Dr. Keren Rice, Dr. Sonya Bird), and all of the Oromo participants who took part in this study. Thank you also to SSHRC for funding the project.

6. References

Audacity Team (2018). *Audacity(R) 2.1.1: Free Audio Editor and Recorder [Computer application]*. URL: https://www.audacityteam.org.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily (2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal". In: *Journal of Memory and Language* 68, pp. 255–278.

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 (1992−2024). *Praat*. http://www.praat.org. University of Amsterdam. URL: ttp://www.fon.hum.uva.nl/praat/.

de Leeuw, Joshua R. (2015). "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser". In: *Behavior Research Methods* 47, pp. 1–12.

Keating, Patricia, Marc Garellek, and Jody Kreiman (2015). "Acoustic properties of different kinds of creaky voice". In: *Proceedings of the 18th International Congress of Phonetic Sciences*. Ed. by The Scottish Consortium for ICPhS 2015. Glasgow, UK: The University of Glasgow.

Lloret, Maria-Rosa (1994). "Implosive Consonants: Their representation and sound change effects". In: *Belgian Journal of Linguistics* 9.1, pp. 59–72.

Owens, Jonathan (1985). *A grammar of Harar Oromo*. Hamburg: Helmut Buske Verlag.

Percival, Maida (2014). "Variation in Ejectives: An Acoustic Study of Stop Contrasts in Eastern Oromo and Dɛlınẹ Slavey". MA thesis. University of Toronto.

— (2018). "An ultrasound and electroglottograph study of voicing in gemination in Eastern Oromo". Paper presented at the 5th NINJAL International Conference on Phonetics and Phonology. National Institute for Japanese Language and Linguistics, Tokyo.

Percival, Maida, Alexei Kochetov, and Yoonjung Kang (2018). "An ultrasound study of gemination in coronal stops in Eastern Oromo". In: *Proceedings of Interspeech 2018*.

R Core Team (2017-2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Acoustic Analysis of Fricatives in Lushootseed

Ted Kye[1]

[1]*University of Washington*

tkye29@uw.edu

Abstract

Although acoustic analysis of fricatives in Salish languages has focused on languages of the Interior Salish branch, there has been little acoustic research on fricatives for Coast Salish. In this study, I examine the acoustic properties of fricatives from archival recordings dating to the 1950s for the Coast Salish language Lushootseed. Spectral moments and intensity of fricatives were examined. The findings suggest that spectral moments differentiated fricative contrasts in Lushootseed by using time-averaging. The findings also provide implications on acoustic-articulatory correlates of fricative contrasts, as well as methodological limitations on the analysis of fricatives from archival recordings dating to the 1950s.

Keywords: Fricatives, spectral moments, time-averaging, intensity, Coast Salish

1. Introduction

The research question for this paper is (1) what the acoustic correlates of fricatives in Lushootseed are, and (2) to what extent can the acoustics of fricatives be analyzed from legacy (archival) recordings dating to the 1950's. 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, 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 these recordings. This study has implications on the production mechanism of fricatives in Lushootseed, methodological issues concerning the acoustic analysis of fricatives using spectral moments from old archival recordings, and implications on the production mechanism of fricatives across Coast Salish.

Lushootseed is a Coast Salish language spoken in the Puget Sound region of the Pacific Northwest. There are two dialects of Lushootseed: Southern Lushootseed and Northern Lushootseed. **Figure 1** illustrates a map of the distribution. There are no fluent native speakers of Lushootseed remaining. For this reason, documentation of the sound patterns of Lushootseed from archival recordings of fluent native speakers is of interest. Like most Salish languages, Lushootseed has a large inventory of consonants, with 37 contrastive consonants (31 are obstruents) and only four contrastive vowels. Although Lushootseed stops and affricates have a three-way laryngeal contrast (voiced, voiceless, and ejective), all fricatives in Lushootseed are voiceless and pulmonic.

There are seven fricatives in Lushootseed: /s ʃ ɬ xʷ χ χʷ h/ ‹s š ɬ xʷ x̌ x̌ʷ h›. Like most Coast Salish languages, dorsal fricatives can have the secondary articulation of labialization. Unlike the uvular place of articulation (which contrasts labialized uvular fricatives from plain uvular fricatives), labiovelar fricatives do not have a plain counterpart. For this reason, it is of interest to characterize the contrast between the dorsal fricatives using acoustic data.

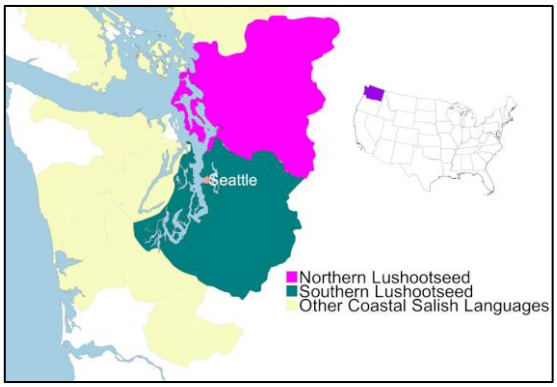


Figure 1: *Map illustrating the distribution of Lushootseed dialects.*

2. Methods

2.1. Recordings

Because there are no fluent native speakers of Lushootseed remaining, several recordings from the archives were examined. These recordings come from the Leon Metcalf collection, which is part of University of Washington's Burke Museum's Special Collections. The recordings were made by the ethnomusicologist Leon Metcalf during the 1950's, where he spent many years traveling across the Puget Sound recording indigenous elder speakers telling stories, myths, oral histories, and private correspondences. These recordings were digitized at 44.1kHz with a 32-bit depth. Eight of these recordings (with a combined length of 57 minutes) were examined. These are recordings of traditional Salish myths and private correspondences (in other words, these are recordings of connected "spontaneous" speech).

2.2. Speakers

Two speakers were examined: Annie Jack and Martha Lamont. Annie Jack is a Southern Lushootseed speaker who was born around the 1870s or 1880s and lived in the Muckleshoot Tribal reservation her entire life. She spoke the Green River, White River, and Duwamish dialects. Her living descendants include Denise Bill (great-granddaughter), Willard Bill, Jr. (great-grandson), Elise Bill-Gerrish (great-great-granddaughter and daughter of Denise Bill), and Justice Bill (great-great-grandson and son of Willard Bill, Jr.). Six recordings of Annie Jack (recordings of traditional Salish myths) were examined. Martha Lamont is a Northern Lushootseed speaker who was born around the 1880s and lived in Tulalip. She spoke the Snohomish dialect. Her living descendants include Hank Williams (grandson), his daughter, and his descendants. Three recordings of Martha Lamont were examined (recordings of private corrrespondences).

2.3. Measurements and sampling procedure

Six of the seven fricatives, /s ʃ ɬ xʷ χ χʷ/, were examined. Each of these fricatives were analyzed with respect to their word position (word-initially, word-medially (intervocalically), and word-finally). Fricatives that occurred in clusters were omitted from the analysis because of the possible effects of C-C coarticulation on the spectral properties of each fricative. A sample size of 40 or more tokens for each fricative were examined.

The software Praat (Boersma & Weenink 2023) was used to analyze the acoustic properties of each fricative. The methods that were used to analyze the acoustic properties of fricatives comes from Shadle (2012, 2023) and Forrest et al. (1988). Multiple Discrete Fourier Transform (DFT) power spectrums were computed for each fricative with a window size of 15ms across the total duration of the fricative. The DFTs from each windowed spectrum were averaged using time-averaging (Shadle 2012; 2023), where the multiple DFTs that were extracted across the total duration of the fricative was averaged through a matrix of intensity and sampling frequencies for each token. Fricatives with a duration less than 56ms were omitted from the analysis to avoid potential overlap in each windowed frame. The Praat script that was used to calculate spectral moments from time-averaged spectrums come from DiCanio (2021).

Following Forrest et al. (1988), spectral moments were calculated from the time-averaged spectrum. These include Center of Gravity (CoG) (also known as the centroid or spectral mean), which is a common method used to measure how high the frequences in a spectrum are (on average) for fricatives (Forrest et al. 1988; Gordon et al. 2002; Shadle 2023); Kurtosis, which measures how narrow the peak is centered around the mean (the higher the Kurtosis, the narrower the peak); Skew, which measures the direction of the skew from the CoG; and Variance, which measures how much the frequencies of the spectrum deviate from the CoG (Forrest et al. 1988). Altogether, these measurements are called "spectral moments" (Forrest et al. 1988; Hargus et al. 2020; Gordon et al. 2002).

The intensity (in dB SPL) of each fricative was extracted from five time points (10%, 30%, 50%, 70%, 90%) across the total duration of each fricative. The intensity of each fricative was analyzed with respect to their word position (word-initially, word-medially, and word-finally). **Figure 2** is an example of a waveform and spectrogram illustrating the time points where intensity was extracted for the fricative [ʃ].

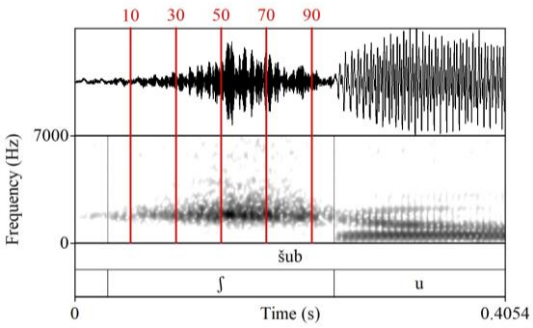


Figure 2: *Illustration of time points where intensity was extracted for the fricative [ʃ].*

2.4. Analysis

Using the statistical software R Studio (2018), the data was fit into a linear mixed effects model using the package *lme4* (Bates et al. 2015), with each spectral moment as dependent variables and each fricative as fixed effects (backwards coded). Speakers and words were treated as random effects, where fricatives were used as random slopes. A *t*-value of 2 or greater was considered significant. The following equation was used to analyze the fricative contrast with respect to their acoustic dimensions:

$$Measurement \sim Fric. + (Fric.|Speaker) + (1|Word) \quad (1)$$

3. Results

Examples of DFT power spectrums for each fricative (extracted from the midpoint of the fricative and taken from word-initial position) can be observed in **Figure 3**. These DFT power spectrums were obtained from a Hamming window of 15ms. The intensity (in dB SPL) of [s] was relatively low compared to the other fricatives. Shadle (2023) observed that [s] has lower intensity levels when produced at low (soft) effort levels. This may suggest that, in connected/spontaneous speech, the production of [s] may have been produced with reduced effort. Another observation worth noting is the maximum frequency (or frequency peak) observed for [xʷ] and [χʷ] when compared with [χ], where the peak for [xʷ] and [χʷ] occurs at a relatively lower frequency than [χ]. This may suggest that the acoustic coupling of lip rounding lowers the frequency of the peak.

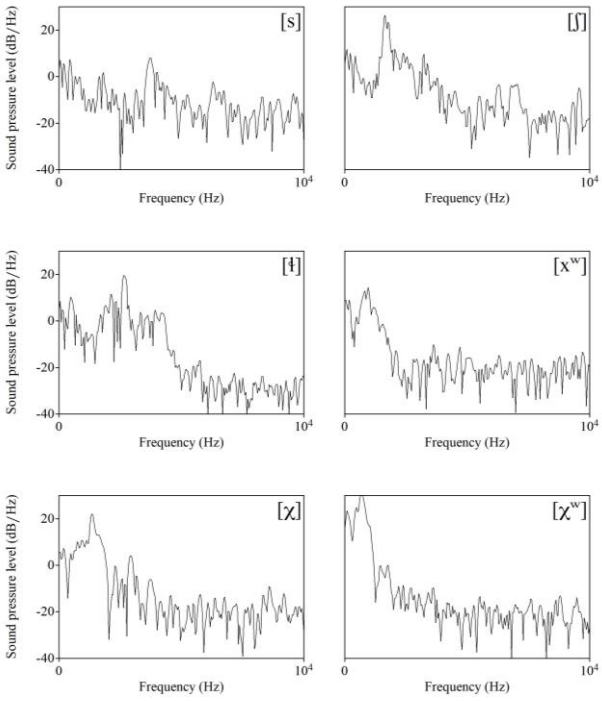


Figure 3: *Examples of DFT power spectrums (15ms Hamming windows) for each fricative in word-initial position.*

As **Figure 4** illustrates, the CoG for [s] was the highest, followed by [ʃ] and [ɬ], followed by [χ], while the labiodorsal fricatives [xʷ] and [χʷ] were the lowest. The current data reveals that the CoG for [s] is significantly greater than [ʃ] (β = 550.539, t = 5.529), [ʃ] not significantly different from [ɬ] (β = -605.807, t = -1.739), [ɬ] significantly greater than [xʷ] (β = 697.771, t = 4.286) and [χ] (β = 491.560, t = 4.850), and [χ] significantly greater than [xʷ] (β = 372.667, t = 2.662) and [χʷ] (β = 365.754, t = 2.896) (i.e., [s] > [ʃ ɬ] > [χ] > [xʷ χʷ]). It should be noted that there was considerable cross-speaker differences in the CoG of [ʃ] and [ɬ], where the CoG of [ʃ] did not significantly differ from

[ɬ] for the speaker Annie Jack but significantly greater for the speaker Martha Lamont.

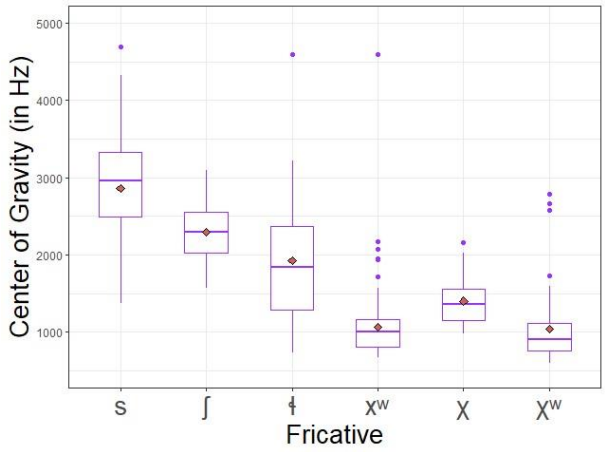


Figure 4: *Distribution of the first spectral moment (i.e., CoG) for each fricative. Means plotted as red diamonds (here and throughout).*

Figure 5 illustrates the distribution of kurtosis for each of the fricatives. As **Figure 5** reveals, [ʃ], [xʷ], and [χʷ] had the highest kurtosis, whereas [s], [ɬ], and [χ] had the lowest kurtosis. The data reveals that [ʃ] has a significantly greater kurtosis than [s] (β = 165.348, t = 5.009) and [ɬ] (β = 110.028, t = 3.621). For the dorsal fricatives, [χ] has a significantly lower kurtosis than [xʷ] (β = -164.529, t = -4.713) and [χʷ] (β = -202.689, t = -5.195). This suggests that the acoustic coupling of labialization yields a narrower spectral peak.

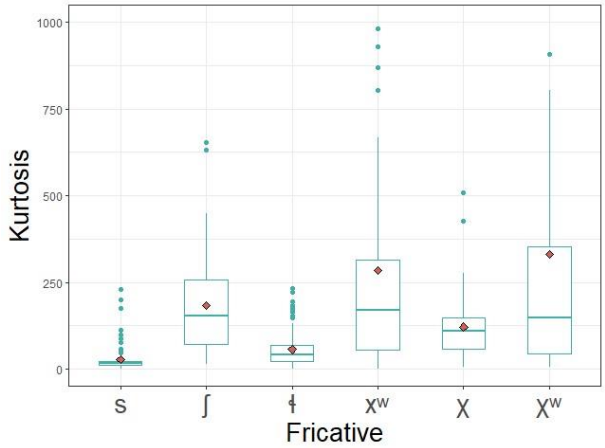


Figure 5: *Distribution of Kurtosis for each fricative.*

Figure 6 illustrates that skew was the greatest for the dorsal fricatives. However, there was also a relatively high skew for the post-alveolar fricative [ʃ], which was commensurate with the uvular fricative [χ]. The data reveals that [ʃ] has a significantly greater skew than [s] (β = 4.366, t = 4.537) but is not significantly different from [ɬ] (β = -2.25, t = -1.622). [xʷ] has a significantly greater skew than [ɬ] (β = 8.222, t = 3.397) and is significantly greater than [χ] (β = 6.0123, t = 2.634). Surprisingly, the labiouvular fricative [χʷ] did not significantly differ in skew from [χ] (β = 4.933, t =1.787). This might be explained by cross-speaker differences in skew. For the speaker Annie Jack, skew for [χ] and [χʷ] did not appear to differ from each other. However, they did appear to differ for the speaker Martha Lamont. This suggests that there are cross-speaker differences in the distribution of skew.

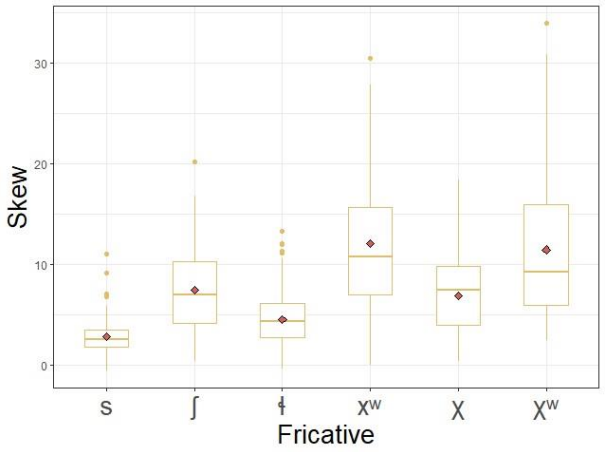


Figure 6: *Distribution of skew for each fricative.*

Figure 7 illustrates the distribution of variance. As **Figure 7** illustrates, the variance for [s] was the highest, whereas [ʃ] was the lowest. The data reveals that [ʃ] has a significantly lower variance than [s] (β = -1367.751, t = -5.195) and [ɬ] (β = -880.233, t = -6.282). [xʷ] had a significantly lower variance than [ɬ] (β = -494.668, t = -2.968). All dorsal fricatives did not significantly differ in variance from each other.

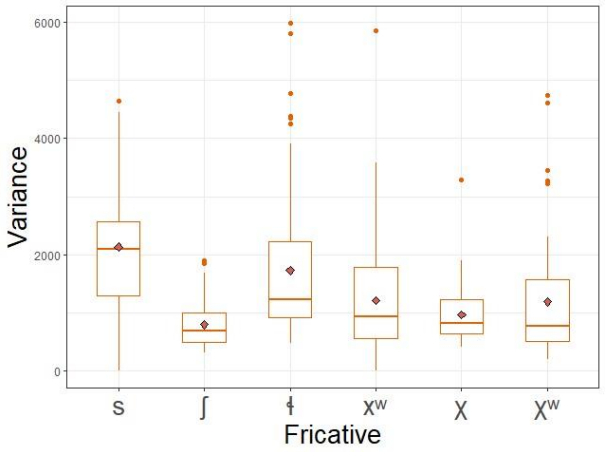


Figure 7: *Distribution of Variance for each fricative.*

Figure 8 illustrates the intensity at five time points (10%, 30%, 50%, 70%, 90%) for each fricative in (1) word-initial position, (2) word-medial position, and (3) word-final position. In all three word-positions, the overall intensity of [s] was the lowest, which may suggest that [s] was produced with less effort level than the other fricatives. In word-initial position, the overall intensity of [ʃ] was the highest. Moreover, unlike the other fricatives (which shows a slight rise in intensity at 90%), [ʃ] is characterized by a slight fall in intensity at the 90% point of the fricative duration. Interestingly, the overall intensity for [χ] was the highest in word-medial position and is characterized by a slight fall in intensity at the 90% point. In word-final position, all fricatives reveal a slight fall in intensity. It should be noted that most of these fricatives in word-final position was followed by a pause, which suggests that energy dissipates towards the end of the frication period when it is followed by silence and not a voiced source.

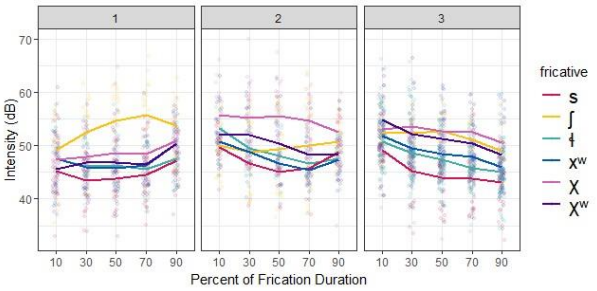


Figure 8: *Intensity at five time-points (10%, 30%, 50%, 70%, 90%) for each fricative. 1 = word-initial, 2 = word-medial, 3 = word-final.*

4. Discussion and conclusion

Spectral moment measurements can differentiate the fricative contrasts in Lushootseed. Where one measure didn't show a contrast, the other did. For example, although [ʃ] and [ɬ] did not differ in CoG, they differed in kurtosis and variance. The labialized dorsal fricatives has higher kurtosis than the plain uvular fricative [χ], which suggests that the acoustic coupling of lip rounding yields a more narrower spectral peak. As expected, the more front the constriction (as is the case of [s]), the higher the CoG. This is due to the smaller cavity in front of the constriction, where the intensity of the noise source becomes prominent at higher frequencies through a shorter channel (Stevens 2000). Because the length of the cavity in front of the constriction for dorsal fricatives is much larger, this would generate a noise source that becomes prominent at lower frequencies when the channel is much longer. It should be noted that there are considerable cross-speaker differences as well. For example, the skew for the plain uvular fricative [χ] did not differ from [χʷ] for the speaker Annie Jack but differed significantly for the speaker Martha Lamont.

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.

There are some noteworthy problems when it comes to measuring spectral moments and intensity for the fricative [s] from these recordings. The CoG for [s] is considerably lower than expected: The mean CoG for [s] was 2858Hz, which is considerably lower than the expected 5~8kHz CoG for that fricative (Gordon et al. 2002; Munson 2001; Forrest et al. 1988). However, it should be noted that there was an upper frequency cutoff from the microphone signal above 6~7kHz. This suggests that absolute measures of CoG for [s] cannot be reliably obtained from these recordings. However, relative differences could nevertheless be obtained for [s] by using time-averaging, where [s] had a significantly greater CoG than [ʃ] (mean CoG = 2295). Praat's default measure of CoG yielded an average CoG of [s] as 1662.38Hz, which was lower that [ʃ]. This suggests that the method of time-averaging is far more accurate at measuring spectral moments from these recordings.

5. Acknowledgements

Many thanks go to the Burke Museum for making these recordings available. I would also like to acknowledge the family members of the speakers in this study: Denise Bill (great-granddaughter of Annie Jack), Willard Bill, Jr. (great-grandson of Annie Jack), Elise Bill-Gerrish (great-great-granddaughter of Annie Jack and daughter of Denise Bill), Justice Bill (great-great-grandson of Annie Jack and son of Willard Bill, Jr.); and Hank Williams (grandson of Martha Lamont), his daughter, and his (and Martha's) descendants. Most of all, I am strongly in debt to the speakers themselves: Annie Jack and Martha Lamont, renowned storyteller's whose legacies will never be forgotten and forever be preserved in these recordings. May the language and spirit of Annie Jack and Martha Lamont continue to live on within their descendants.

6. References

Bates, Douglas, Martin Mächler, Ben Bolker, & Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4." In *Journal of Statistical Software, 67*(1), 1-48.

Boersma, Paul & David Weenink (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.50, retrieved 20 June 2021 from http://www.praat.org/.

DiCanio, Christian (2021). Spectral moments of fricative spectra script in Praat. Retrieved from https://www.acsu.buffalo.edu/~cdicanio/scripts.

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.

Hargus, Sharon, Gina-Anne Levow, Richard Wright (2020). Acoustic Characteristics of Deg Xinag Fricatives. In *University of Washington Working Papers in Linguistics*, 1-53.

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).

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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.

Stevens, Kenneth N. (2000). *Acoustic Phonetics*. MIT Press.

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

Abstract

The study investigates the interplay between speech motor control and cognitive executive dysfunction by looking at inter-articulatory coordination patterns between consonants and vowels in the production of syllables with high (CCV) and low (CV) complexity. Kinematic speech data (EMA) of 25 people with Parkinson's disease (PwPD) and 25 healthy controls (HC) were recorded. Further, the influence of levodopa on syllable coordination as well as the relationship to cognitive executive dysfunctions were tested. Results showed preserved articulatory coordination on the level of intra-syllabic coordination. On the intra-gestural level, consonantal and vocalic movements were prolonged in the PD group and positively affected by the intake of levodopa. For the PD group, a correlation between the shift pattern of the second consonant and scores on the executive function test is found, indicating that executive dysfunctions possibly give rise to changes in articulatory timing patterns.

Keywords: *Parkinson's disease, speech motor control, articulatory coordination, executive functions*

1. Introduction

Speech production requires the control over motor processes and cognitive functions, 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 (Ziegler & Vogel 2010). 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). Articulatory movements are therefore affected in various ways: speech movements are smaller in amplitude, slower and consequently longer in duration, and articulatory coordination is compromised when comparing it to healthy control speakers (Yunusova et al. 2008, Ziegler & Vogel 2010). PD also affects cognitive processes, including working memory, attention, executive control, and visuospatial domains (Aarsland et al. 2021). Executive functions, one of the most frequently impaired functions in PD (Kalbe et al. 2016), play a crucial role in orchestrating cognitive processes. They are thought to be an umbrella term comprising, amongst others, set-shifting abilities (Kudlicka et al. 2011) - the ability to switch between different tasks or mental sets. Executive functions/set-shifting skills can be assessed with the Trail Making Test (TMT). There is only a very limited number of studies on the kinematics of speech in PwPD in general and while a first step has been made to relate acoustic speech parameters to cognitive dysfunction (Thies et al. 2020), its relationship to articulation is yet to be explored.

In the present study, we therefore investigate the interplay between speech motor control and cognitive dysfunction by examining kinematics of syllable coordination patterns in syllables with branching onsets (/pl/) in the production of PwPD and HC. We also explore the role of dopaminergic substitution in form of levodopa, the most common and effective form of treatment for PD, on these timing patterns by additionally comparing medication OFF (med-OFF) and ON (med-ON) status within the PD group. We then correlate the articulatory findings with cognitive scores of the TMT. We tested the following assumptions for the PD group when being compared to the HC group: (i) We expect the PD group to produce deviant articulatory timing patterns in syllables with high complexity. (ii) We expect a positive effect of levodopa on the inter-gestural coordination. (iii) We expect that articulatory changes in the timing of syllables with high complexity correlate with lower performance scores on the executive functions test.

2. Methods

The analysed data were collected as part of a larger study by Thies (2023) conducted at the Department of Neurology of the University Hospital in Cologne. We here investigate a subset of the data that has not been looked at so far.

2.1. Participants

25 PwPD (5 female, 20 male) aged between 40 and 77 (mean age = 60 years) and 25 age- and sex-matched HC participated in the study. Four HC had to be excluded from the analysis due to issues with the sensor tracking leading to inaccurate data trajectories or due to incorrect articulation of the target words. Of the HC included in the analysis, 3 were female and 19 male (mean age = 61 years). All participants were native speakers of German and underwent a screening process to rule out the presence of dementia or depression. Motor functions for both groups were assessed using part III of the Unified Parkinson's Disease Rating Scale (UPDRS-III). The PwPD were diagnosed between 1 to 20 years (mean = 8 years) prior to study inclusion and were recorded in both med-OFF and med-ON conditions. Med-OFF involved withdrawing PD medication for at least 12 hours, while med-ON entailed the intake of a predetermined standardized levodopa dosage of 200 mg.

2.2. Neuropsychological assessment

All participants underwent a neuropsychological assessment. The TMT was administered to assess executive functions. It consists of two parts: In part A, participants are asked to connect a sequence of consecutive numbers from 1 to 25; In part B, participants have to connect a sequence of numbers (1 to 13) and letters (A to L), alternating between the two (i.e., 1-A-2-B etc.). The time needed to complete TMT-A serves as an indicator for processing speed, and the time score of TMT-B allows for drawing conclusions on mental flexibility which is related to set-shifting. Additionally, the TMT difference score (B-A) and ratio score (B/A) were calculated as they are said to control for influences of motor control and other non-set-shifting elements, thereby emphasizing executive functions (Muir et al. 2015). One person in the HC group did not complete the TMT. The neuropsychological assessment for the PwPD was only carried out in med-ON condition.

2.3. Speech recording and speech material

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 (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.

2.4. Speech data annotation and measurements

Speech data were processed in the EMU-webAPP of the EMU-SDMS environment (Winkelmann et al. 2017). On the acoustic level, we calculated segment durations of the first stressed syllables. 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). Therefore, target positions of the articulators for consonants (C_1, C_2) and vowels (V) in the first stressed syllables were identified in the vertical plane using zero-crossings in the respective velocity trace. To measure the overlap, latencies between the maximum target positions of C_1, C_2, and the C-centre (midpoint between two Cs) to the V were computed. Consonantal shifts were calculated by comparing the latencies in CV and CCV syllables: 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/) with C_1C_2V (/pli/) (latency should decrease from CV to CCV).

2.5. Statistical analysis

The data was analysed using the statistical computing software R (version 4.3.3; R Core Team, 2024). To test differences in acoustic segment durations and articulatory timing patterns between syllable structures (CV vs. CCV) and between groups/conditions (HC vs. med-OFF, HC vs. med-ON, med-OFF vs. med-ON), linear mixed effect models were conducted. Syllable structure and group/medication condition were set as predictor variables, and random intercepts for intra-speaker variability were included. For the correlation analysis, the difference in shift of C_1 and C_2 between CV and CCV syllables was correlated with the different TMT scores (A, B, B-A, B/A) of the HC group and the PwPD (med-ON). The data were tested for normal distribution in which case the Pearson method was used for the correlation analysis. Otherwise, the Spearman method was applied. Based on the first round of results, some additional correlation analyses were performed: latencies/syllable durations ~ TMT scores, UPDRS ~ syllable durations/C_2 shift/TMT scores. Interaction effects between TMT and UPDRS scores on C_2 shifts were also tested with linear models.

3. Results

3.1. Acoustic data

The mean acoustic segment durations for C_1 /p/, C_2 /l/, V /i/ and for the entire syllable are reported in Table 1. Results show that durations of C_1 /p/ do neither differ between syllable structures nor between groups/conditions ($p > .05$ across all comparisons). Durations of C_2 /l/ are shorter in CCV compared to CV syllables ($p < .001$ across all comparisons, mean difference = -56.2 ms). The comparison between groups/conditions shows longer C_2 durations for med-OFF compared to the HC ($p = .012$, mean difference = 21.3 ms). The durations decrease from med-OFF to med-ON, eliminating group differences between med-ON and HC ($p > .05$). The durations of the vowel /i/ do not differ between syllable structures across all groups/conditions ($p > .05$). However, med-OFF presents with longer V durations both compared to HC ($p = .010$, mean difference = 28.5 ms) and to med-ON ($p < .001$, mean difference = 16.1 ms). The durations decrease from med-OFF to med-ON, eliminating group differences between med-ON and HC ($p > .05$). Thus, durations of CCV syllables are longer compared to CV syllables and longer durations of C_2 and V in med-OFF lead to longer syllable durations in this condition (Table 1).

Table 1: *Means and sd of acoustic durations in ms specified by group/condition and by syllable structure.*

		/p/	/l/	/i/	syllable
HC	C_1V	197 (64)	—	121 (37)	318 (90)
	C_2V	—	96 (34)	133 (37)	230 (64)
	C_1C_2V	205 (78)	57 (24)	117 (36)	379 (110)
med-ON	C_1V	188 (45)	—	123 (31)	324 (67)
	C_2V	—	121 (68)	144 (43)	265 (94)
	C_1C_2V	186 (62)	55 (20)	136 (40)	364 (86)
med-OFF	C_1V	200 (44)	—	150 (47)	350 (77)
	C_2V	—	130 (57)	166 (48)	296 (89)
	C_1C_2V	197 (68)	67 (23)	142 (54)	406 (106)

3.2. Articulatory data

The mean articulatory latencies between C_1 /p/, C_2 /l/, and C-centre and the vocalic anchor respectively are reported in Table 2. The latencies of /p/ to /i/ increase from CV to CCV across all groups/conditions ($p < .001$, mean difference = 68.6 ms). When comparing groups/conditions, latencies of C_1 to V only differ between med-OFF and med-ON, i.e., they are longer in med-OFF ($p < .001$, mean difference = 20 ms). The latencies of C_2 /l/ to V differ slightly between CV and CCV syllables in the two PD conditions only, i.e., they are longer in CV compared to CCV ($p = .026$, mean difference = 9.73 ms). When comparing groups/conditions, the C_2 latencies are longer in med-OFF both compared to HC ($p = .016$, mean difference = 30.8 ms) and compared to med-ON ($p < .001$, mean difference = 15.8 ms). The latencies decrease from med-OFF to med-ON eliminating group differences between med-ON and HC ($p > .05$). The med-OFF/med-ON effect is further reflected in the shortening of the latency between the C-centre and the vocalic anchor ($p = .001$, mean difference = -22.2 ms).

Table 2: *Means and sd of articulatory latencies in ms specified by group/condition and by syllable structure.*

		C_1 to V /p/ → /i/	C_2 to V /l/ → /i/	C-centre → /i/
HC	C_1V	177 (63)	—	—
	C_2V	—	119 (51)	—
	C_1C_2V	254 (75)	119 (47)	186 (52)

med-ON	C_1V	188 (52)	—	—
	C_2V	—	141 (42)	—
	C_1C_2V	243 (53)	130 (38)	187 (43)
med-OFF	C_1V	201 (50)	—	—
	C_2V	—	158 (55)	—
	C_1C_2V	274 (68)	147 (52)	211 (56)

The shift pattern for the complex onset /pl/ is visualised in Figure 1. It shows the leftward (negative values) and rightward (positive values) shifts of C_1 and C_2 respectively. For all groups/conditions we find a clear leftward shift of C_1. Looking at the C_2 shift, a small rightward shift for med-ON and med-OFF, and a small leftward shift for HC become apparent.

Figure 1: *Shift patterns of C_1 (in red) and C_2 (in blue) from CV to CCV. Shift direction: < 0 to the left, > 0 to the right.*

3.3. Executive functions and correlations

The mean TMT performance scores of the two groups are shown in Table 3. TMTA differed between the two groups with the PD group taking longer to complete the test ($p = .042$, mean difference = 8.28 ms). The same group difference can be observed for TMTB and the derived scores (Table 3).

Table 3: *Means and sd of TMT scores in s.*

	TMTA	TMTB	TMTB-A	TMTB/A
HC	30.3 (10.2)	78.1 (31.3)	47.8 (28.4)	2.67 (0.9)
med-ON	38.8 (14.9)	94.1 (61.6)	55.3 (54.4)	2.42 (1.03)

Correlations between shift patterns of C_1 and C_2 and the different TMT scores were first assessed across the two groups and then for each group individually. For the across groups analysis, there was a single correlation between C_1 shift and TMTA ($p = .005$, $r_S = .411$). When looking at the two groups individually, visual inspection revealed two outlier points (> 2 sd) in the PD group. To make sure the correlations are not driven by these outliers, analyses were performed excluding the two speakers. The results are shown in Table 4, revealing that C_2 shift corelates with all executive function scores. This correlation, exemplified by C_2 shift ~ TMTB-A, is shown in Figure 2: More extreme rightward shifts of C_2 are associated with higher TMTB-A scores. No such correlations were found for the HC group.

Table 4: *P-values and correlation coefficients between C_2 shift and TMT scores.*

	C_2 shift ~		
	TMTB	TMTB-A	TMTB/A
med-ON	$p = .044$ $r_P = .423$	$p = .008$ $r_P = .540$	$p = .010$ $r_S = .526$

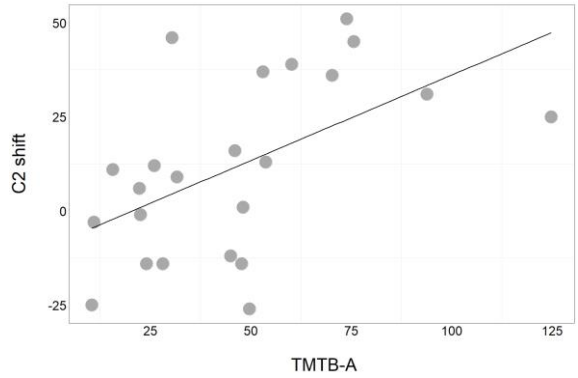


Figure 2: *Correlation of C_2 shift with TMT difference score (TMTB-A). Shift direction: < 0 to the left, > 0 to the right.*

The acoustic and articulatory analyses showed a trend pattern between groups with durations of C_2 and V as well as all latency measures being longest in med-OFF > med-ON > HC (with the exception of C_1 latency in CCV for med-ON). At the same time, the intra-syllable coordination patterns were stable across groups/conditions. Therefore, we tested the relationship between changes on the temporal level and TMT performance. In a first step, we correlated the latencies of C_1 and C_2 in CCV to syllable duration. Results show that both latencies correlate with syllable durations (latency C_1 ~ syllable duration: $p < .001$, $r_S = .872$; latency C_2 ~ syllable duration: $p < .001$, $r_S = .555$). We then tested the correlation between syllable duration and the different TMT scores across groups as well as for HC and PD separately. No correlations were found for any of the measures.

Finally, we tested whether the disease severity might have affected our measures. To validate our results, we tested for correlations with the UPDRS scores. Syllable durations did not correlate with the UPDRS scores for all groups/conditions. No correlation between UPDRS scores and C_2 shift could be found across groups/conditions and for individual groups/conditions. Looking at PD in med-ON only, UPDRS scores did correlate with the TMT pure scores (TMTA: $p = .011$, $r_S = .497$; TMTB: $p = .032$, $r_S = .430$). However, no correlation between UPDRS scores and TMT derived scores (difference/ratio) could be found. A slight interaction effect between TMTB and UPDRS scores was found for the C_2 shift in med-ON condition ($p = .042$). Visual analysis of this interaction showed that with increasing UPDRS scores, the correlation between TMTB and C_2 shift also increases.

4. Discussion and conclusion

In line with prior research, the articulatory results reveal a non-symmetrical timing pattern for the complex onset coordination /pl/ for neurotypical speakers of German. While C_1 /p/ presents with a leftward shift, C_2 /l/ does not shift considerably towards the following V from CV to CCV. Instead, the acoustic C_2 segment was shortened in CCV due to coarticulatory effects of the jaw, lips, and tongue in terms of compensatory shortening

(e.g. Pouplier 2012, Mücke et al. 2020). The same non-symmetrical timing pattern was observed in PwPD for complex syllable organization, even in a poor motor status, i.e., without medication. It is noteworthy, that while there is a small leftwards shift of C_2 present in the HC group, both PD conditions show a slight C_2 shift to the right. Both phenomena are not surprising and have been reported for /pl/ in German in previous research (Mücke et al. 2020). 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 C_2 and V: In the med-OFF condition, PwPD produced longer consonantal and vocalic movements, and these durational changes on the intra-gestural level led to longer latencies between Cs and Vs on the inter-gestural level. A general trend of med-OFF > med-ON > HC emerged, where group differences between med-ON and HC are often eliminated. This underlines a beneficial effect of levodopa on speech planning abilities, which has been shown before (e.g. Thies et al. 2021).

Turning towards the neuropsychological test scores, group differences were observed for all TMT scores with PwPD presenting with lower performance scores than the HC. However, we want to point towards a limitation of this study as TMT scores for the PD group were only obtained in the med-ON condition. Some studies show that set-shifting skills are likely to improve under levodopa which might lead to even more significant differences between HC and PwPD (Gul & Yousaf 2022). Nonetheless, we did find a correlation between the C_2 shift and all TMT scores involving part B of the test within the PD med-ON group. This relationship between timing patterns and executive functions, particularly set-shifting, lets us assume that some PwPD change their articulatory timing patterns as C_2 tends to shift more to the right when there is a decline in set-shifting abilities, indicating the possibility of a less efficient/deviant timing. The general tendency of stable coordination patterns that are scaled in time that we observed for the PwPD did not correlate with the executive function scores which suggests that impaired executive function skills might be the cause of articulatory timing changes. A further, in detail analysis of PwPD who present with larger rightward shifts of C_2 and lower scores in TMT performance and their clinical characteristics might paint a clearer picture of the processes at play.

The correlation of the motor scores with TMTA within the PD med-ON group are an indicator that processing speed declines as a function of disease severity. It explains the group differences found between HC and PD med-ON regarding their performance on the TMTA. As UPDRS scores correlated with TMTB, their interaction effect on C_2 shift was investigated. It was found that increased disease severity, measured by UPDRS scores, likely reinforces the negative effect executive dysfunction has on the shift pattern of C_2. However, future research might want to investigate whether this interpretation holds when further disease severity and cognitive measures are considered, or whether what we observed is rather a parallel decline of speech motor control and executive functioning.

All things considered, this study contributes to the understanding of the interplay between speech motor control and cognitive executive functioning. Its findings indicate that speech therapy concepts might need to be adapted to include cognitive training, which in turn will have a positive effect on speech symptoms of PD. In the future, it might be worthwhile to replicate the present study including assessment of the TMT for the med-OFF condition to be able to assess the influence of levodopa on executive functioning and correlate timing patterns with executive function skills both on and off medication. We would expect an even stronger correlation for the PD med-OFF group. Also, including consonant cluster types other than /pl/ in the analysis would deepen our understanding of how articulatory shift patterns and executive function skills correlate in neurotypical and impaired speech. There are compensatory shortening mechanisms in German /pl/, that might have increased the variability in our data.

5. Acknowledgements

This work was supported by the German Research Foundation (DFG) as part of the SFB1252 “Prominence in Language” (Project-ID 281511265).

6. References

Aarsland, D., Batzu, L., Halliday, G.M., Geurtsen, G.J., Ballard, C., Chaudhuri, K.R. & Weintraub, D. (2021). Parkinson disease-associated cognitive impairment. *Nature Reviews Disease Primers 7*(1): 47. DOI: 10.1038/s41572-021-00280-3.

Gul, A., & Yousaf, J. (2022). Efficacy of L-dopa in Treatment of Aggression, Frontal Lobe Cognitive Functioning and Task Switching Deficits in Parkinson’s Disease Patients. *Pakistan Armed Forces Medical Journal*, *72*(SUPPL-2), 132–135. DOI: 10.51253/pafmj.v72iSUPPL-2.3354.

Kudlicka A., Clare L. & Hindle J.V. (2011). Executive functions in Parkinson's disease: systematic review and meta-analysis. *Movement Disorders 26*(13), 2305-2315. DOI: 10.1002/mds.23868.

Muir, R.T., Lam, B., Honjo, K., et al. (2015). Trail Making Test Elucidates Neural Substrates of Specific Poststroke Executive Dysfunctions. *Stroke 46*(10), 2755-2761. DOI: 10.1161/STROKEAHA.115.009936.

Mücke, D., Hermes, A. & Tilsen, S. (2020). Incongruencies between phonological theory and phonetic measurement. Phonology 37(1), 133-170. DOI: 10.1017/S0952675720000068.

Pouplier, M. (2012). The gestural approach to syllable structure: Universal, language-and cluster-specific aspects. In Fuchs, S., Weirich, M., Pape, D. & Perrier, P. (eds.) *Speech planning and dynamics*. Frankfurt am Main: Peter Lang. 63–96.

R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

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. (2023). *Tongue Body Kinematics in Parkinson’s Disease: Effects of Levodopa and Deep Brain Stimulation*. Berlin: Peter Lang Verlag.

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.

Thies, T., Mücke, D., Lowit, A., Kalbe, E., Steffen, J. & Barbe, M.T., (2020). Prominence marking in parkinsonian speech and its correlation with motor performance and cognitive abilities. *Neuropsychologia 137*: 107306. DOI: 10.1016/j.neuropsychologia.2019.107306.

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: 10.1044/1092-4388(2008/043).

Ziegler, W. & Vogel, M. (2010). *Dysarthrie: Verstehen, Untersuchen, Behandeln*. Stuttgart: Georg Thieme Verlag.

Why do palatographic data have to be taken seriously?

Yury Makarov

Institute of Linguistics, RAS; Vinogradov Russian Language Institute, RAS (Russia);
University of Cambridge (United Kingdom)

im562@cam.ac.uk

Abstract

In this paper, I argue that palatography is a highly informative tool despite its apparent simplicity. The paper begins with an overview of the evolution of palatography, highlighting its intricate detail and variations. Following this, various techniques of palatography are introduced, along with associated challenges, which partly question established views. Subsequently, I suggest linguistic applications of palatography, focusing on its potential to describe place of articulation and inform phonological typology. Moreover, I propose that the inclusion of palatographic data in linguistic accounts can explain contact-induced changes in phonological systems and intra- and inter-speaker articulatory variation. This proposition is supported by a preliminary panel study of a Shughni speaker who demonstrated changes in articulation over one year.

Keywords*: speech production, instrumental phonetics, history of phonetics, phonological typology, palatography.*

1. Introduction

In modern-day phonetic/phonological research preference is often given to sophisticated instrumental techniques, often requiring substantial funding, advanced technical skills and/or a lot of equipment. Examples include pneumotachography (measuring nasal/oral airflows; Barry & Kuenzel, 1975; Dewhurst, 2023), electroglottography, laryngeal endoscopy (vocal fold activity; Herbst, 2020), ultrasound/MRI imaging (various aspects of articulation, including tongue and larynx movement; Gick, 2002; Hudu, 2014; Mielke et al., 2017; Takano & Honda, 2007), electroencephalography (brain activity; Mai et al., 2022). For additional methods, refer to (Gick et al., 2013). While these methods undoubtedly further our understanding of language mechanisms, most of them are confined to a laboratory setting.[1] This presents a problem for field linguists, who often give up on instrumental phonetic research because of the costs and complexity involved, restricting the phonetic component of their research agenda to collecting acoustic data.

However, certain techniques, despite being cheap, easy to use, and informative, often go unnoticed by linguists. This paper focuses on palatography and argues for its necessity in any fieldwork project concerned with language documentation. I begin by describing the history and different varieties of this technique; this is followed by the discussion of practical aspects of doing palatography in the field and interpreting the obtained data. Finally, I turn to the question of how palatographic data can be applied to describing the sounds of a language, changes in articulation, and phonological typology.

2. The evolution of palatography

The first instance of using a substance applied to the mouth to investigate the physiology of speech is believed to belong to James Oakley Coles, a London dentist. The procedure he invented around the 1870s involved spreading a sticky substance over the soft and hard palate as well as the upper teeth and, after articulating 'a letter [name],' describing where the mixture has been removed (Abercrombie, 1957). This method, later called *direct palatography*, differs from that introduced by the New York dental surgeon Norman William Kingsley in 1879. Kingsley's version of palatography involved the use of an artificial palate (Abercrombie, 1957; Ashby, 2016, p. 58).

In the following years, somewhat of a boom in the application of palatography to phonetic research happened. Phoneticians from different countries published articulatory studies based on palatographic evidence (see (Gósy, 2023) on Hungarian phoneticians and (Gordina, 2006) on Vasily Bogoroditsky's work on Russian) and even devised special apparatuses (as James Anthony at the Edinburgh Phonetics laboratory). As in many other cases (e.g., see (Makarov, 2024) w.r.t. the concept of reduction), Eduard Sievers's 'Grundzüge der Phonetik' (Principles of Phonetics) played a significant role in establishing palatography as a mainstream technique (Ashby, 2016, p. 59).

Classic palatography only gives a static representation of the tongue–palate contact during the production of a sound. Furthermore, if several sounds in the stimulus involve contact with the roof of the mouth, the palatogram will be difficult to interpret because of the superimposition of a series of traces. Despite these limitations, palatography was used for studying coarticulation already at the beginning of the 20th century (Hardcastle, 1981, p. 59). In (Ladefoged, 1957, p. 768) it is even stated that 'in most phonetic investigations it is advisable to make palatograms showing the effect of pronouncing whole words… [it] is preferable to the artificial procedure of attempting to obtain a record of an isolated speech sound.' The palatograms of *key* and *coo* are then given as references illustrating a shift in the place of articulation of the velar plosive depending on the following vowel.

In the early 1960s, a new version of palatography, electropalatography (EPG), was developed, now employing an artificial palate with metal electrodes associated with certain anatomical landmarks (Hardcastle, 1972; Kuzmin, 1962). With no paint involved, a dynamic study of tongue–palate contact became possible (for details see Hardcastle & Gibbon, 2014).

[1] It is worth mentioning that field linguists are trying to adjust some of these techniques to the fieldwork setting, particularly ultrasound imaging (Gick, 2002; Timkin, 2022).

Later, when computer-based display systems became available[2], EPG found multiple applications in speech therapy.

Nowadays, in spite of some attempts to reintroduce classic direct palatography into active use (most notably Ladefoged, 2003) as the most convenient means of collecting data on place of articulation, the technique seems to receive limited attention from both phoneticians (as an old-time, unsophisticated research method) and field linguists (not having enough motivation to apply the technique). While there are a few recent studies using data from direct palatography (e.g., Chen & Guo, 2022; Chirkova et al., 2015; Coretta et al., 2023), there is no impression that palatographic evidence has become an essential part of every language description (cf. its absence in plenty of JIPA's illustrations).

3. Techniques of palatography

Since the invention of palatography in the 19th century, several variants of the technique have been used for studying slightly different aspects of articulation. Palatography is called **direct** when no artificial palate is used (opposite: **indirect**). Another distinction lies between its **static** and **dynamic** variants; the former does not give information about the production of every segment in a stimulus (which is usually a word composed of several sounds), yielding a snapshot of all lingual gestures involved. On the contrary, dynamic palatography, also known as **electropalatography**, traces the articulation of every segment involving tongue–palate contact (Hardcastle, 1972; Hardcastle & Gibbon, 2014). Finally, much variation is related to what the paint is applied to. In classic palatography marks are made on the roof of the mouth and teeth by the tongue covered with non-toxic paint. The resulting pictures of the passive articulators are called palatograms. Conversely, if painted is the roof of the mouth, and marks on the tongue are photographed, such technique is called **linguography**[3] and obtained pictures are linguograms (Gick et al., 2013, p. 181). However, if the regions of the roof of the mouth (or the artificial palate) where the marking medium was wiped away are inspected instead, this is still palatography (Ladefoged, 1957, p. 764).

In all kinds of palatography except for EPG some kind of marking medium is used. In the earlier versions of the technique, this substance could include a range of ingredients including meal, mucilage, ink, chalk and even alcohol (Abercrombie, 1957; Gósy, 2023, p. 684; Witting, 1953). However, in modern versions, a mixture of edible cooking oil and powdered charcoal is used as it is non-toxic and almost tasteless (Anderson, 2008, p. 5; Ladefoged, 2003, p. 38). Sometimes instead of the oily mixture a black powder made of charcoal and drinking chocolate is sprayed (Abercrombie, 1957, p. 23; Ladefoged, 2003, p. 45).

4. Issues in palatography

4.1. Analysing the photographs

Since palatography is used primarily for identifying the place of articulation (or sometimes the part of the tongue involved) the palatograms/linguograms need to be mapped onto some kind of articulatory categories (e.g., dental, alveolar, palato-alveolar; apical, laminal; etc.). Concerns were raised regarding the loss of information about palatal morphology. As Ladefoged (1957) puts it, 'A view of the palate from a point at right angles to the dental… preserves the ratio between the length and the width of the palate only at the expense of giving an inadequate impression of the depth of the palate. As a result, palatograms often fail to convey important information concerning the shape and depth of the palatal cavity, and the position and slope of the alveolar ridge.' To solve this problem, Ladefoged suggests that a cast of the mouth be made and sawn along the mid-line. However, it is rather unclear whether this 'important information' is linguistically relevant and essential for drawing conclusions concerning place features.[4] There is only a limited number of hypotheses on how palatal morphology could influence phonology (e.g., Makarov, 2022, p. 161; Moisik & Dediu, 2020) and therefore it seems that for linguistic research the loss of the third dimension in a palatogram is negligible.[5]

A useful practice is using zones of the roof of the mouth as reference points. In spite of differences in dentition, it is usually possible to determine the frontmost contact in the palatogram. While distinguishing between dentals and alveolars is quite straightforward, further articulations can be assessed based on the horizontal lines corresponding to specific teeth as suggested by (Firth, 1948), whose system was successfully used, for example, in (Kim, 2001; Makarov, 2025). Ladefoged (1957, p. 772) criticised Firth's using the teeth as reference points on the basis of (a) 'insufficient correlation between the positions of specific teeth and the positions of anatomical features… which are important in determining the acoustic quality of a speech sound' and (b) 'several teeth may be missing, and there may or may not be gaps between the teeth which remain; sometimes the teeth… overlap; and nearly always the posterior molars are not far enough back to provide adequate reference points on the soft palate.' As for (a), it is doubtful that assigning a place label has to be in any way affected by acoustics (see Section 5.1 on dentals vs. alveolars, which are difficult to distinguish by ear); challenges evoked by dentition in (b) seem to be mitigated if Firth's zones are perceived not as absolute but rather relative reference points. It is usually possible to reconstruct the zones based on their expected widths, even if the teeth are absent or displaced.

[2] The first techniques of EPG required a high-speed camera to photograph the read-out panel rendering contact areas with a number of circular spots of light corresponding to the electrodes in the artificial palate (Hardcastle, 1972).

[3] Technically, it is still palatography, cf. its treatment in (Ladefoged, 2003; Witting, 1953). The term *linguography* emerged as an attempt to clarify what part of the mouth is painted first, though it seems quite unnecessary as eventually both the palate and the tongue get marked. Moreover, when palatography was invented (see Section 2), a sticky substance was spread over the palate, just like the paint in linguography. It is also noteworthy that attempts were made to call palatograms 'linguagrams' instead (Abercrombie, 1957, p. 22).

[4] As (Witting, 1953, p. 60) puts it, 'there is a **theoretical** [highlighted by me. — Y. M.] possibility of a correlation between palatal anatomy and articulations taking place in that region.'

[5] Especially for a field linguist, for whom making impressions means more weight in the backpack and more excuses to obtain an informant's consent.

4.2. Number of speakers and choosing the technique

One of the most important issues to consider when undertaking palatographic research is the number of speakers sufficient for drawing conclusions. As put by Ladefoged (2003, p. 31), 'a sufficient number of speakers [is required] to make sure that you are describing properties of the language, and not just the personal characteristics of one or two people.' The problem is that *sufficient* is dependent on many factors, one of the most crucial being how many people there are to work with. Most of the palatographic studies seem to rely on data from only a few speakers, from one to four. Hypothetically, it may be enough to capture a possible articulation but surely does not suffice to detect variation and determine the relative frequency of the variants. Such a small number of participants is especially upsetting when the language in question is not minor but has millions of speakers. For example, in (Kim, 2001), the study of Korean sounds is based on data from only four subjects while the number of Korean native speakers is ca. 80 million. Although getting subjects' consent can be difficult in some cultures as the procedure involves physical interaction with the mouth, it is not impossible to get a higher number of subjects even for minor languages. For instance, in (Makarov, 2025) data from seven speakers of Shughni are analysed, and Shughni is spoken only by ca. 100,000 people in the Pamir Mountains. It is necessary for the researcher to clearly explain the technique and demonstrate its safety, which usually helps in obtaining consent.

The choice of the specific kind of palatography is also related to the number of subjects one can get. It is obviously not possible for an average field linguist to perform EPG or even indirect palatography since making an artificial palate is too resource-intensive.

4.3. Choice of stimuli

Another important issue is the choice of words to be investigated. As was discussed in Section 2, pronouncing separate sounds should be avoided as it is likely to evoke unnatural articulations. An ideal token for the basic study of sound production (not coarticulation) should have only one lingual consonant paired with an open vowel. Having more than one lingual gesture requiring contact with the roof of the mouth will mar the palatogram and hence decrease its reliability. Open vowels like [a] and labial consonants are good supplements to the target sound.

4.4. Synchronisation with audio recording

In the history of palatography, many attempts were made to synchronise the palatographic procedure with audio recording (e.g. Witting, 1953). While it can be useful in theory, in reality, it seems to be redundant. To make sure that the studied utterance is natural the researcher has to supervise the procedure on-site and ask to repeat it in case of failure. Using the resulting recording for acoustic analysis will not be particularly fruitful as it is only one repetition without any carrier phrase (so it will not even be possible to perform statistical analysis). Moreover, if indirect palatography is used, pronunciation will inevitably be altered because of the artificial palate. The only reasonable application of synchronised audio is for dynamic palatography as it can help inspect particular stages of sound articulation.

5. Linguistic application of palatography

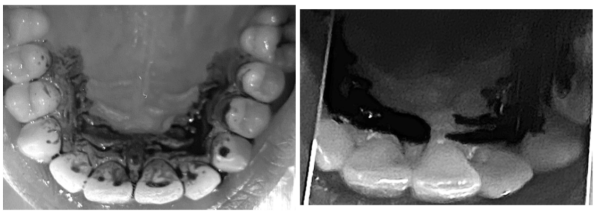
In the following sections, several applications of palatography will be discussed. All of them are linguistically relevant and useful for both language documentation and theoretical matters.

5.1. Classification of coronal sounds

The attribution of coronals to dentals or alveolars often lacks any clear explanation, not to mention instrumental evidence, though a palatographic study offers an easy solution to the problem. 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.

Typologically, the dental–alveolar phonemic 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 support from 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. The study of Shughni coronals (Makarov, 2025) 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.

5.2. Beyond the front of the mouth: 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, 2025): 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.

6. Conclusions

In this paper, I discussed the history, different techniques and applications of palatography. Despite it may seem unsophisticated, palatography has abundant detail and can be used for studying a variety of topics, not limited to articulations of a particular language.

7. Acknowledgements

I am grateful to St Edmund's College (Cambridge) for partially supporting my participation in ISSP 2024.

8. References

Abercrombie, D. (1957). Direct Palatography. *Zeitschrift Für Phonetik*, *10*, 21–25. https://doi.org/10.1524/stuf.1957.10.14.21

Anderson, V. B. (2008). Static Palatography for Language Fieldwork. *Language Documentation & Conservation*, *2*(1), 1–27.

Ashby, M. (2016). *Experimental phonetics in Britain, 1890-1940* [PhD Thesis]. University of Oxford.

Barry, W., & Kuenzel, H. (1975). Co-articulatory airflow characteristics of intervocalic voiceless plosives. *Journal of Phonetics*, *3*(4), 263–281. https://doi.org/10.1016/S0095-4470(19)31434-2

Chen, Y., & Guo, L. (2022). Zhushan Mandarin. *Journal of the International Phonetic Association*, *52*(2), 309–327. https://doi.org/10.1017/S0025100320000183

Chirkova, K., Wang, D., Chen, Y., Amelot, A., & Kocjančič Antolík, T. (2015). Ersu. *Journal of the International Phonetic Association*, *45*(2), 187–211. https://doi.org/10.1017/S0025100314000437

Coretta, S., Riverin-Coutlée, J., Kapia, E., & Nichols, S. (2023). Northern Tosk Albanian. *Journal of the International Phonetic Association*, *53*(3), 1122–1144. https://doi.org/10.1017/S0025100322000044

Dewhurst, M. (2023). Enrichment of Sociolinguistic Nasality Research with Phonetic Data: Methodological Considerations. *Modern Languages Open*, *2023*(1), 7. https://doi.org/10.3828/mlo.v0i0.453

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

Firth, J. R. (1948). Word-Palatograms and Articulation. *Bulletin of the School of Oriental and African Studies, University of London*, *12*(3/4), 857–864.

Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, *32*(2), 113–121. https://doi.org/10.1017/S0025100302001007

Gick, B., Wilson, I., & Derrick, D. (2013). *Articulatory phonetics*. Wiley-Blackwell.

Gordina, M. V. (2006). *Istorija foneticheskikh issledovanij (ot antichnosti do vozniknovenija fonologicheskoj teorii) [The history of phonetic research (from antiquity to the emergence of phonological theory]*. Filologicheskij fakul'tet SPbGU.

Gósy, M. (2023). On the history of palatography in Hungarian phonetics. *Journal of the International Phonetic Association*, *53*(3), 682–693. https://doi.org/10.1017/S0025100321000293

Hardcastle, W. J. (1972). The Use of Electropalatography in Phonetic Research. *Phonetica*, *25*(4), 197–215. https://doi.org/10.1159/000259382

Hardcastle, W. J. (1981). Experimental Studies in Lingual Coarticulation. In R. E. Asher & E. J. A. Henderson (Eds.), *Towards a history of phonetics: Papers contributed in honour of David Abercrombie* (pp. 50–66). Edinburgh University Press.

Hardcastle, W. J., & Gibbon, F. E. (2014). Electropalatography as a Research and Clinical Tool. In W. J. Hardcastle & J. Mackenzie Beck (Eds.), *A Figure of Speech* (pp. 39–60). Routledge. https://doi.org/10.4324/9781410611888

Herbst, C. T. (2020). Electroglottography – An Update. *Journal of Voice*, *34*(4), 503–526. https://doi.org/10.1016/j.jvoice.2018.12.014

Hudu, F. (2014). [ATR] feature involves a distinct tongue root articulation: Evidence from ultrasound imaging. *Lingua*, *143*, 36–51. https://doi.org/10.1016/j.lingua.2013.12.009

Karamshoev, D. (1963). *Badzhuvskij dialekt shugnanskogo jazyka [Bajuwi dialect of Shughni]*. Izdatel'stvo AN Tadzhykskoj SSR.

Kim, H. (2001). The place of articulation of the Korean plain affricate in intervocalic position: An articulatory and acoustic study. *Journal of the International Phonetic Association*, *31*(2), 229–257. https://doi.org/10.1017/S0025100301002055

Kuzmin, Y. I. (1962). Mobile palatography as a tool for acoustic study of speech sounds. *Proceedings of the 4th International Congress of Acoustics*.

Ladefoged, P. (1957). Use of Palatography. *Journal of Speech and Hearing Disorders*, *22*(5), 764–774. https://doi.org/10.1044/jshd.2205.764

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Blackwell Pub.

Mai, A., Riès, S., Ben-Haim, S., Shih, J., & Gentner, T. (2022). Phonological contrasts are maintained despite neutralization: An intracranial EEG study. *Proceedings of the Annual Meetings on Phonology*, *9*. https://doi.org/10.3765/amp.v9i0.5197

Makarov, Y. (2022). [Review of:] B. Sands (ed.). Click consonants. Leiden; Boston: Brill, 2020. *Voprosy Jazykoznanija*, *2*, 157–162. https://doi.org/10.31857/0373-658X.2022.2.157-162

Makarov, Y. (2024). Reduction: A Brief History of the Concept (18th–20th Centuries). *Russian Speech = Russkaya Rech'*.

Makarov, Y. (2025). *Shughni consonants in production: A palatographic study.* To appear in*: Advances in Iranian linguistics. Vol. III.* John Benjamins Publishing Company.

Mielke, J., Carignan, C., & Thomas, E. R. (2017). The articulatory dynamics of pre-velar and pre-nasal /æ/-raising in English: An ultrasound study. *The Journal of the Acoustical Society of America*, *142*(1), 332–349. https://doi.org/10.1121/1.4991348

Moisik, S., & Dediu, D. (2020). The ArtiVarK Click Study: Documenting Click Production and Substitution Strategies by Learners in a Large Phonetic Training and Vocal Tract Imaging Study. In B. Sands (Ed.), *Click Consonants* (pp. 384–417). BRILL. https://doi.org/10.1163/9789004424357_013

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.

Takano, S., & Honda, K. (2007). An MRI analysis of the extrinsic tongue muscles during vowel production. *Speech Communication*, *49*(1), 49–58. https://doi.org/10.1016/j.specom.2006.09.004

Timkin, T. V. (2022). First syllable vowels in Surgut Khanty according to the ultrasonography data. *Siberian Journal of Philology*, *3*, 196–211. https://doi.org/10.17223/18137083/80/16

Wilkins, D. P. (1989). *Mparntwe Arrernte (Aranda): Studies in the structure and semantics of grammar* [Doctor of Philosophy]. The Australian National University.

Witting, C. (1953). New techniques of palatography. *Studia Linguistica*, *7*(1–2), 54–68. https://doi.org/10.1111/j.1467-9582.1953.tb00490.x

Influence of stress and sequence position on vowel sandhi in Brazilian Portuguese

João Paulo Moraes Lima dos Santos[1,2]

[1]*University of Salamanca, Spain*
[2]*Federal Institute of Sertão Pernambucano, Brazil*
joaopaulomls@gmail.com / joao.paulo@usal.es

Abstract

This paper presents an analysis of how the variables of word stress, intonational phrase stress, and the position of the vowel sequence within the phrase affect the production of vowel sequences across word boundaries, generating some external sandhi phenomena. Based on semi-spontaneous data from Brazilian Portuguese speakers, the study suggests that the stress pattern of a word can influence the occurrence of sandhi phenomena or the maintenance of hiatus. However, this also depends on whether the sequence carries the main stress of the intonational phrase, as in "no próximo ano," and if it is located at the rightmost phrase boundary, as in "isso que é." It is observed that the probability of a sandhi phenomenon occurring is higher in contexts where at least the first vowel is unstressed, when neither of the vowels receives the main stress of the intonational phrase, and when they are not at the phrase boundary.

Keywords: Vowel sequences, stress, sequence position, Brazilian Portuguese

1. 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 some sandhi phenomena. This reduction can manifest as a monophthong (e.g., "*camisa usada*" pronounced as [kɐ̃.mi.zu.za.da]) or a diphthong ([kɐ̃.mi.zaʊ̯.za.da]). Stress appears to be a crucial factor influencing the execution of these processes (Abaurre, 1996; Bisol, 2003, 2013, Silva, 2012).

In the context of stress, these prior studies 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 processes. Notably, some studies also consider the main stressed accent of the intonational phrase as a variable influencing these processes (Abaurre, 1996; Bisol, 2002, 2013; Tenani, 2004, Oliveira & Santos, 2018). In the context of prosody, we refer to Nespor & Vogel (1986), who classify the intonational phrase (hereinafter 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) 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.

Tenani (2004), based on the prosodic phonology of Nespor & Vogel (1986), identifies the intonational phrase as the prominent domain for the contraction of vowel sequences. Thus, in a phrase like "*a aluna aceitou o convite*" the fusion of the emphasized vowels is easily applied in Brazilian Portuguese. On the other hand, Tenani verifies the blocking of fusion of identical vowels when the subject is viewed as an intonational phrase independent of the verb. According to her, the blocking occurs when the subject belongs to one phrase and the verb to another phrase. For example, the author observes the blocking of fusion in the sequence /aa/ emphasized in the phrase "*a aluna, após o exame, foi para a discoteca*" Likewise, Ludwig-Gayer & Dias (2017), in a study on the process of coalescence of identical vowels ('*degeminação*') in the variety of the city of Salvador, Brazil, verify the importance of the intonational phrase for the application of vowel fusion. Additionally, the authors find the atonicity of the vowels as one of the factors favoring the process. Oliveira & Santos (2018) describe the transition from the primary stressed accent at the word level, such as "*isso*" ['i.sʊ] in Portuguese, to secondary at the IP level, as in "*mas é isso aqui*" [i.sʊ̯a.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 this sense, in addition to stress, we have also examined the behavior of sequences based on the primary stressed accent at the IP level along with the position of the vowel sequence. We also investigate whether the position of the sequence in the intonational phrase (within or at the limit of the IP) affects the vowel sequences. The investigation explores how these factors interact and contribute to the observed sandhi and hiatus patterns in our data.

Therefore, this study aims to analyze the effects of stress and vowel sequence position on sandhi processes and hiatus maintenance. Our hypothesis is that there are more possibilities of contraction when both vowels are unstressed. However, when there is a context with at least one stressed vowel, the contraction is conditioned by two factors: if the stressed vowel of the word does not receive the primary stress of the IP and is not located at the rightmost boundary of the phrase.

2. Method

2.1. Participants

The description and analyses are based on semi-spontaneous data obtained through interviews with ten native speakers from the city of Recife, Brazil. A total of 1,509 vowel sequences across word boundaries were recorded. All participants signed a document of informed consent to participate in the study voluntarily. Table 1 lists the labeling details, sex of the informant, and the quantity of vowel sequence productions analyzed in this study.

Table 1: *Labeling, sex, and number of productions of vowel sequences for each participant in the research.*

Participant	Sex	*n*
Rec1	Female	176
Rec2	Female	155
Rec3	Female	170
Rec4	Male	180
Rec5	Male	150
Rec6	Male	158
Rec7	Male	101
Rec8	Female	159
Rec9	Male	147
Rec10	Female	113
TOTAL		1509

2.2. Materials

Recordings were made using a computer, a unidirectional microphone, and a Scarlett 2.0 audio interface. The data were captured and stored in Audacity software with a sampling frequency set at 44,100 Hz. Interviews were conducted in offices known for their favorable acoustics, situated within either the Faculty of Philology or the Center for Brazilian Studies at the University of Salamanca, Spain.

2.3. Acoustic analysis of sandhi/hiatus production

Acoustic analysis was performed using the freely available software PRAAT (Boersma & Weenink, 2019), version 6.0.53. The formant points were extracted based on the Praat scripts by Barrientos (2019a; 2019b), adapted for this study with the compilation of three monophthong points and eleven diphthong and hiatus points.

The extraction of formant values for monophthongs was performed using points relative to 20%, 50%, and 80% of the spectral space of the vowel (labeled as f1.2, f1.5, and f1.8, respectively).

Distinctions between diphthong and hiatus sequences were established based on the parameter of the presence or absence of formant stability in the segments (Barbosa & Madureira, 2015). For diphthongs and hiatuses, eleven points were labeled from the beginning to the end of the spectral space of the sequence (labeled as f1.0 for the start of the sequence, followed by f1.1, f1.2, and so forth, up to f1.10 for the end of the sequence). This approach allowed for a more detailed observation of the degree of stability or lack thereof in the vowels that transition into glides.

The method used for extracting formant values was cepstral analysis. Unlike LPC analysis, which has limitations for sounds with antiformants, cepstral analysis can be applied to any type of sound. Therefore, this technique appears to be the most suitable for our data analysis, considering potential interference from adjacent nasal segments and the analysis of nasal vowels and nasalization in Brazilian Portuguese sequences. Additionally, for more precise results, formant values in the spectral region overlapping with the cepstrum were considered, as described in Barbosa & Madureira (2015).

Figure 1: *Sound wave, spectrogram, and TextGrid of the monophthong production in the phrase 'd***o** **a***mbiente' (participant Rec4).*

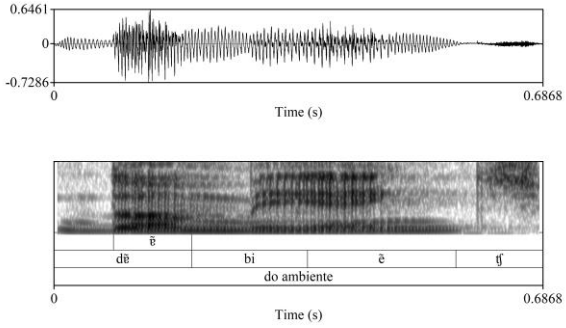
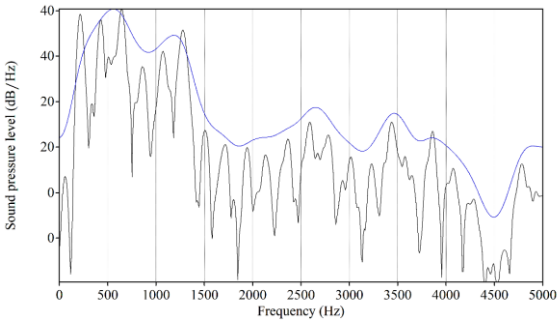


Figure 2: *Spectrum (black line) and cepstrum (blue line) of the realization of [ɐ̃] at approximately 50% point of the vowel in the phrase 'do ambiente'. The approximate values of F1 and F2 at the peaks of the cepstrum were 560Hz and 1186Hz, respectively.*



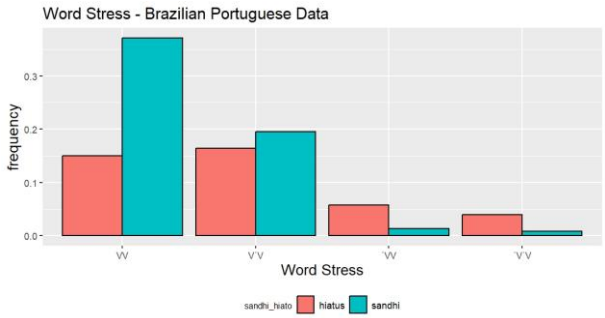
2.4. Statistical analysis

In RStudio, a mixed logistic regression model (Faraway, 2016) was utilized for statistical analysis. This model incorporated the following variables: (1) the type of production as response variable; (2) the accent of vowels within sequences (whether stressed or unstressed), intonational phrase stress (indicating primary stressed accent), and sequence position (whether it occurs at a phrase boundary) as fixed effects; and (3) speaker identity as a random effect. The glmer function from the lme4 package (Bates et al., 2015) was employed in R for model implementation. Additionally, to enhance comprehension of the model's data, log-odds results were transformed into probabilities using the invlogit function from the scales package (Wickham & Seidel, 2022), and tab_model from the sjPlot package (Lüdecke et al., 2021).

3. Results

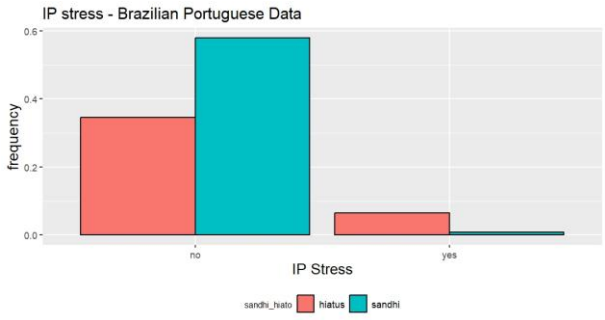
The descriptive analysis of the Brazilian Portuguese data shows a higher prevalence of sandhi in contexts where i) both vowels are unstressed and ii) when the first vowel is unstressed and the second is stressed. In the stressed/unstressed and stressed/stressed contexts, hiatus is maintained in most of the data. The following graph (Figure 3) illustrates the relative frequency of sandhi/hiatus production according to the stress of the vowels at word boundaries.

Figure 3: *Relative frequency of Brazilian Portuguese data grouped by word stress.*



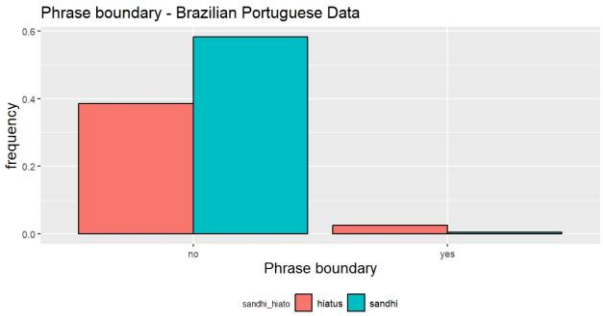
Word stress may affect the occurrence of a sandhi process or the maintenance of the hiatus, but that will also depend on whether the vowel sequence carries the primary stressed accent of the intonational phrase (as, for example, in "*então no próxim**o a**no*") and whether it is at the limit of the phrase at the right (as in "*pois isso qu**e é***"). Considering whether any of the vowels receive the primary stress of the intonational phrase, the data show a higher number of hiatus productions in that context. When none of the vowels carry the primary stress, sandhi is applied in most of the data. The relative frequency for the presence/absence of the primary stress can be verified in Figure 4:

Figure 4: *Relative frequency of Brazilian Portuguese data grouped by the presence/absence of the primary stress of the intonational phrase.*



The descriptive analysis also identifies a greater number of hiatuses when the sequence is located at the right edge of the phrase, as observed in the example "*mas no geral o brasileir**o é***". In contrast, in contexts where the sequence is not located at the right edge of the phrase, sandhi production significantly exceeds hiatus maintenance. The bar chart in Figure 5 provides the description of the data according to the sequence's location in the intonational phrase.

Figure 5: *Relative frequency of Brazilian Portuguese data grouped by the position of the sequence in the intonational phrase (IP).*



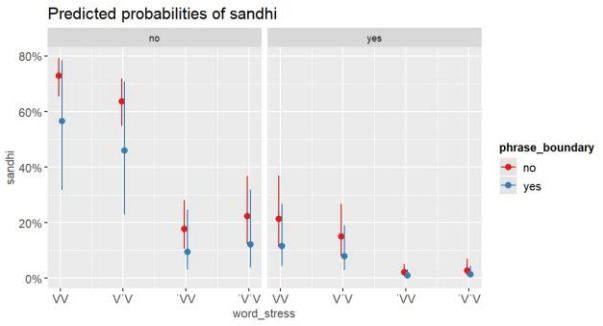
The results of the regression model indicate that the tendency for a sandhi to occur is higher in contexts in which at least the first vowel is unstressed, both vowels do not receive the primary stressed accent of the IP, and both vowels are not at the limit of the phrase. The intercept with a positive value (Table 2) 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 phrase, 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 phrase 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.

Table 2: *Mixed logistic regression model for the analysis of stress and vowel sequence position.*

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**
phraseboundary[yes]	-0.73	-1.71 – 0.25	0.145

Figure 6: *Predicted probabilities of sandhi in Brazilian Portuguese data, based on word stress, intonational phrase stress, and phrase boundary. Confidence intervals are also indicated (α = 0.05).*



Finally, ˈVV and ˈVˈV vowel environments have a much higher probability of maintaining the hiatus, regardless of sequence position ('phraseboundary' variable) or the primary stressed accent ('IP' variable). In both contexts, the probabilities and confidence intervals exceed 50%. What is also observed is a greater preference for maintaining the hiatus when these sequences receive the main stressed accent and are placed at the limit of the phrase.

4. Discussion and conclusion

In addition to stress, our study highlights the significance of the primary stressed accent within the intonational phrase (IP) and the position of the vowel sequence within the utterance in influencing sandhi processes in Brazilian Portuguese. Our results underscore the nuanced nature of sandhi phenomena, particularly regarding the resistance to contraction observed in sequences containing stressed vowels. This resistance appears to be contingent upon the alignment of stressed syllables within the sequence with the primary stress of the IP. When this alignment is absent, contraction processes at the vowel boundary become more prevalent, suggesting a dynamic interaction between stress patterns and phonological processes in shaping speech patterns.

In conclusion, 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 (2002) and Tenani (2004). However, our results reveal an interesting twist: sequences with at least one stressed vowel can show 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 doesn't match the primary stress of the IP, 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.

5. Acknowledgements

The author would like to thank Dr. Fernando Sánchez Miret and Dr. Miguel Oliveira Jr. for their valuable comments and suggestions on this paper. Additionally, thanks to the 10 Brazilian Portuguese speakers for their availability to contribute to this research.

6. References

Abaurre, M. B. M. (1996). Acento frasal e processos fonológicos segmentais. Letras de Hoje, v. 31, n. 2, p. 41-50.

Barbosa, P., & Madureira, S. (2015). *Manual de fonética acústica experimental: aplicações a dados do português*. Cortez.

Barrientos, F. (2019a). *Praat scripting I: Basic Operations [Tutorial]*. Https://Www.Fernandabarrientos.Cl/Praat1.Pdf.

Barrientos, F. (2019b). *Praat scripting II: Perceptual experiments [Tutorial]*. Https://Www.Fernandabarrientos.Cl/Praat2.Pdf.

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: Maria Bernadete Abaurre (Org.). *Gramática do português culto falado no Brasil: a construção fonológica da palavra*. 1. ed. São Paulo: Contexto, v. VII, 53-74.

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: Bisol, Leda. Introdução a estudos de fonologia do português brasileiro. Porto Alegre: EDIPUCRS.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC Press Taylor & Francis Group.

Lüdecke, D. (2018). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.11. https://strengejacke.github.io/sjPlot/.

Ludwig-Gayer, J. E., & Dias, L. B. (2017). A degeminação na fala popular de Salvador. *Estudos Linguísticos e Literários*, *57*, 186–206.

Nespor, M., & Vogel, I. (1986). Prosodic phonology. Dordrecht: 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.

Silva, V. M. de O. (2012). Sândi vocálico externo na fala do Rio de Janeiro. *Anais Do IV Seminário Internacional de Fonologia*.

Tenani, L. E. (2002). Domínios prosódicos no português do Brasil: implicações para a prosódia e para a aplicação os processos fonológicos. (PhD Thesis). Universdade Estadual de Campinas, Brazil.

Tenani, L. E. (2004). O bloqueio do sândi vocálico em PB e PE: evidências da frase fonológica. Organon, v. 18, n. 36, p. 17-29.

Wickham, H., & Seidel, D. (2022). *scales: Scale functions for visualization*. https://scales.r-lib.org.

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.unicampr@gmail.com, silveira.gustavocampos@gmail.com

Abstract

This pilot study explores how the mandible times its vocalic opening movements with acoustic vowel onsets (AVO). 3-D Electromagnetic Articulographic along with acoustic data were recorded for three North American English speakers producing utterances with one word in the utterance produced with contrastive emphasis. Analysis of the acoustic and articulatory data used a newly-implemented Praat algorithm. The timing of four mandible movement landmarks in the articulatory data were measured relative to AVO in the acoustic signal: (1) minimum value of acceleration curve (minAcc); (2) minimum value of velocity curve (minVel); (3) maximum value of acceleration curve (maxAcc) and (4) minimum mandible position (maxDisp). The results showed that minVel closely matched AVO timing for two of the three speakers, while maxAcc showed the closest timing to AVO for the other speaker; interestingly, the emphasized word starting with an initial voiceless aspirated consonant [k^h] showed a significantly closer timing to AVO with maxAcc than minVel.

Keywords*: mandible landmarks, velocity, acceleration, displacement, emphasis*

1. Introduction

Work by a number of researchers, (e.g., Erickson et al. 2012; Erickson et al. 2020; Erickson and Niebuhr 2023; Erickson et al. in press; Svensson Lundmark 2023; Svensson Lundmark and Erickson 2023; Svensson Lundmark and Erickson 2024; 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 in the F1-F2 regions of syllable onsets to show that the 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 (2024), 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 broadband spectrograms; and how this timing is affected by changes in syllable prominence, given that the jaw lowers more with increased prominence, (e.g., de Jong 1995; Erickson et al. 2012; Harrington et al. 2000).

2. 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. Utterance (1) is the neutral, for comparing with each of the utterances (2-5). 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 analyzed using newly-implemented Praat algorithms (Barbosa 2023; Silveira 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 (minAcc) associated with mandible beginning to open for vowel, (2) minimum value of velocity curve (minVel) while mandible is opening, (3) maximum value of acceleration curve (maxAcc) for when mandible was open and (4) minimum mandible position (maxDisp) 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 (see Fig.1 for a graphic representation, which also shows maxVel, for mandible raising for the coda consonant).

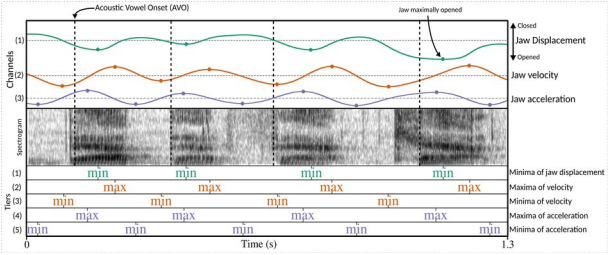


Figure 1: *Mandible vertical movement signal (Channel 1) and its first and second derivatives (Channels 2 and 3), synchronized with broadband spectrogram (0 to 5kHz), of the phrase "bat that fat CAT" by a male speaker (A00).*

3. Results

3.1. Timing of mandible movement with AVO

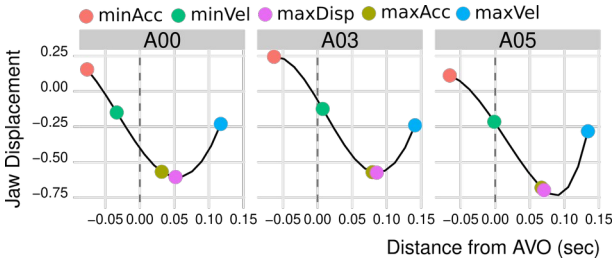


Figure 2: *Schematic drawing of the overall pattern of mandible movement for all syllables in all utterances for each of the three speakers, A00, A03, A05, from left to right. The x-axis indicates the point in time of each mandible landmark (minAcc,maxAcc, minVel, maxVel, maxDisp) relative to the AVO; the y-axis, the amount of mandible lowering at that point in time. For each speaker and each landmark, the mean distance from the AVO was computed, and then a spline interpolation was used to connect the points.*

Figure 2 illustrates the relative timing of various mandible landmarks (velocity, acceleration, and maximum lowering points) relative to the acoustic onset of the vowel (AVO), the 0.0 point on the x-axis, for each pattern. In the schematic drawing, first we see minimum acceleration (red dot), which is the point where the mandible starts to open for the vowel; it represents the height of the mandible when the mandible is closed for production of the initial consonant constriction at the beginning of the syllable. Next, we see minimum velocity (green dot), the point where the mandible is opening fastest as it opens for the vowel nucleus. Just before the mandible reaches its maximum opening for the vowel nucleus (maxDisp) represented by the magenta dot, we see maximum acceleration (olive green dot). Finally, as the mandible approaches its highest point for closure for the syllable coda constriction, we see maximum velocity (maxVel) (blue dot, not analyzed here). Figure 2 suggests that the timing of AVO with minVel for overall mean values is independent of (a) amount of mandible lowering for vowel nucleus and (b) amount of mandible raising for initial consonant constriction. That is, Speaker A05 shows the largest amount of mandible lowering while Speaker A03 shows the highest mandible raised position but both speakers show close timing of minVel with AVO. A note about the terms of maximum and minimum — here we are using the mathematical maxima and minima; in terms of physiology, however, the start of the mandible opening, here referred to as minAcc (red dot), is physiologically maximum acceleration as the mandible accelerates to open; minVel (green dot) is physiologically maximum velocity, where the mandible is moving its fastest in opening; maxAcc (olive green dot) is, in physiologically terms, the point where the mandible decelerates before it reaches maximum opening; and maxVel (blue dot) is when the mandible slows down before reaching closure for constriction of the coda consonant. The actual overall time values of the mandible landmarks from AVO are shown in Table 1. Negative numbers indicate the time in sec before AVO while positive numbers indicate time in sec after AVO.

***Table 1**: Descriptive statistics of the distances of mandible landmarks (s) from AVO for three speakers*

measure	mean	sd	median
maxDisp	0.054	0.056	0.064
maxVel	0.117	0.034	0.118
minVel	-0.011	0.046	-0.006
maxAcc	0.049	0.047	0.059
minAcc	-0.061	0.071	-0.074

Looking at the average of all speakers, minVel of the mandible opening is the measure that occurs closest to AVO 64% of the time. However, as shown in Figure 2, for two of the speakers (A05, A03) minVel occurs at or very close to AVO, but for the third speaker (A00), minVel occurs a certain distance before AVO. For speakers A05 and A03, minVel is the measure closest to AVO 80% of the time, but for speaker A00, the measure that occurs closest to AVO is maxAcc (olive green dot)—the point where the mandible decelerates as it approaches maximal opening for the vowel.

3.2 Effect of emphasis on timing of measured landmarks with AVO

In terms of overall results, regardless of whether the word was emphasized, minVel was the point closest to AVO, with emphasized words having a higher frequency of closeness to AVO (71%) than not-emphasized words (64%). A Kruskal-Wallis test, one for each type of measurement of distance from AVO, was done to compare emphasized vs not-emphasized words. The results show that minVel is not significantly different for emphasized vs not-emphasized words, although both maxAcc and maxDisp significantly vary (p~0.00. As shown in Table 2, for the emphasized words, maxDisp (the maximum amount of mandible lowering as represented by the magenta dot in Figure 2) occurs significantly further to the right of AVO compared to not-emphasized words; consequently, maxAcc (olive green dot in Figure 2), the point where the mandible decelerates before reaching maximum opening, also occurs significantly further from AVO for emphasized words.

***Table 2**: Descriptive statistics of the distances of mandible landmarks from the AVO in seconds according to emphasis. Pairs in bold are significantly different.*

measure type	emph	mean	sd	median
maxDisp	**emphasized**	**0.081**	**0.05**	**0.088**
maxDisp	**not emphasized**	**0.057**	**0.052**	**0.063**
minVel	emphasized	-0.013	0.044	0.002
minVel	not emphasized	-0.012	0.041	-0.011
maxAcc	**emphasized**	**0.067**	**0.05**	**0.079**
maxAcc	**not emphasized**	**0.051**	**0.042**	**0.056**
minAcc	emphasized	-0.069	0.069	-0.07
minAcc	not emphasized	-0.062	0.07	-0.075

3.2 Effect of emphasized word on timing of landmarks with AVO

An interesting finding, however, is that timing of mandible landmarks varies depending on the word that is emphasized. For all emphasized words, except for *CAT*, minVel is the measure that occurs closest to AVO at a frequency of 80% to 100% of the time; however, for emphasized *CAT*, maxAcc(olive green dot in Figure 2) is the landmark that occurs closest to AVO 65% of the time, at 0.021 s BEFORE the AVO (see Table 3). MinVel for *CAT* occurs 0.071 s AFTER the AVO.

***Table 3**: Descriptive statistics of the distances of mandible landmarks from the AVO in seconds according to emphasized word.*

measure	word	mean	sd	median
maxDisp	THAT	0.095	0.022	0.1
maxDisp	**CAT**	**0.065**	**0.03**	**0.065**
maxDisp	BAT	0.117	0.027	0.119
maxDisp	FAT	0.05	0.077	0.083
minVel	THAT	0.008	0.027	0.023
minVel	**CAT**	**-0.071**	**0.036**	**-0.086**
minVel	BAT	0.007	0.014	0.007
minVel	FAT	0.012	0.02	0.013
maxAcc	THAT	0.086	0.025	0.098
maxAcc	**CAT**	**0.021**	**0.048**	**0.016**
maxAcc	BAT	0.097	0.023	0.096
maxAcc	FAT	0.07	0.057	0.079
minAcc	THAT	-0.071	0.037	-0.054
minAcc	CAT	-0.056	0.123	-0.133
minAcc	BAT	-0.086	0.028	-0.076
minAcc	FAT	-0.065	0.025	-0.067

Table 3 shows how mandible movement landmarks for the emphasized words vary depending on the initial consonant onset of the emphasized word. Especially we see this for *CAT*, where minVel occurs BEFORE AVO, with maxAcc the landmark closest to AVO. MaxDisp for *CAT* and *FAT* tend to occur closer to AVO compared to that for *BAT* and *THAT*; maxVel for *CAT* occurs further after AVO than the other three words. The highlighted values in Table 3 show the significantly different values for α=0.05 based on Wilcoxon tests. It is interesting that *CAT*, which begins with a voiceless aspirated [k^h], is significantly different from the two emphasized words that begin with voiced consonants, i.e., [b] and [ð], in terms of maxDisp, minVel and maxAcc but not significantly different in terms of minAcc (the onset of mandible opening). As for *FAT* (which starts with a voiceless fricative) it is significantly different from *CAT* in terms of minVel and maxAcc, but not the other measures.

3.3. Effect of initial consonant and emphasis condition on timing of landmarks relative to AVO

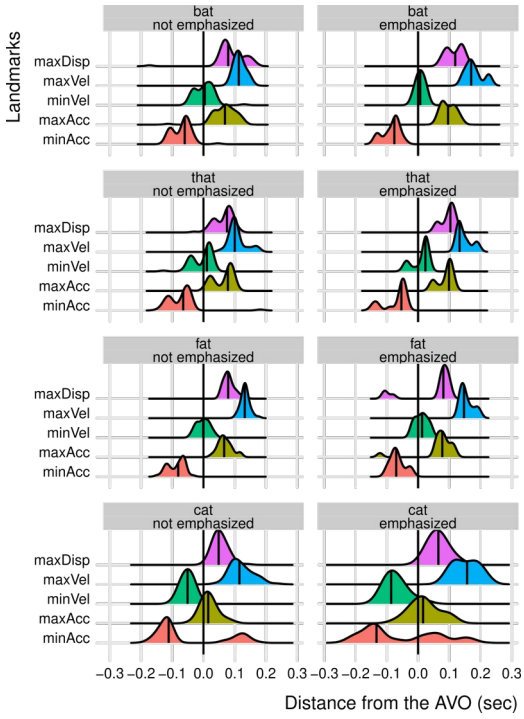


Figure 3: *Timing of mandible landmarks with respect to AVO in seconds as a function of initial consonant of word and emphasis condition. "True" indicates the word was emphasized; "false" indicates the word was not emphasized. The 0-point on the x-axis indicates AVO.*

Figure 3 shows the timing of mandible landmarks with respect to AVO as a function of whether the word was emphasized or not, as well as to the initial consonant of the word. The top panel shows *bat* vs *BAT*, the next panel shows *that* vs *THAT*, then *fat* vs *FAT*, and bottom panel, *cat* vs *CAT*. MaxDisp tends to occur later for emphasized words than for non-emphasized words, but only the *BAT-bat* and *THAT-that* pair show a significant difference. MaxAcc occurs significantly further from AVO for the *FAT-fat* and *CAT-cat* pair. MinVel occurs closer to AVO, either just at AVO or shortly after, but only the *BAT-bat* pair is significantly different. Interestingly, minVel for both *CAT* and *cat* occurs BEFORE AVO, with minVel for *CAT* occurring even further to the left of AVO. For all cases the alpha level was 0.05.

3.4. Effect of amplitude of mandible displacement at AVO on timing of mandible landmarks relative to AVO

Figure 4 shows a linear relation between distance of minVel from AVO and the amplitude of mandible displacement at AVO. A linear regression with amplitude of jaw displacement at the AVO as the dependent variable and the distances of the four landmarks (minVel, maxVel, minAcc, maxAcc) from the AVO as the independent variables, indicated that all variables are statistically significant, as shown in Table 4.

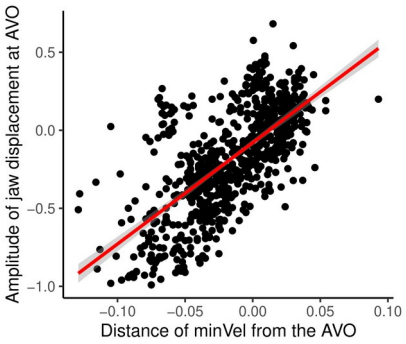


Figure 4: *Distance from minVel from AVO is shown on the x-axis; amplitude of mandible displacement at AVO is shown on the y-axis.*

Table 4: *Descriptive statistics of the distances of extreme points from the AVO according to word (only emphasized occurrences).*

	Estimate	Std. Error	t value	p-value
(Intercept)	-0.20223	0.01712	-11.814	< 0.001 ***
minVel	3.20913	0.26450	12.133	< 0.001 ***
maxVel	-0.28238	0.09102	-3.102	0.002 **
minAcc	0.59643	0.1424	4.186	< 0.001 ***
maxAcc	2.19553	0.25586	8.581	< 0.001 ***

4. Discussion and conclusion

The timing of four mandible movement landmarks in the articulatory data were measured relative to AVO in the acoustic signal: (1) minimum value of acceleration curve (minAcc); (2) minimum value of velocity curve (minVel); (3) maximum value of acceleration curve (maxAcc) and (4) minimum mandible position (maxDisp). The results showed that minVel showed the closest timing to AVO for two of the three speakers, while maxAcc showed the closest timing for the other speaker. Emphasis significantly affected the timing of maxDisp as well as maxAcc with AVO, but did not significantly affect the timing of minVel with AVO. An interesting finding was an effect of the initial consonant of the emphasized word on the timing of mandible landmarks with AVO; specifically, the emphasized word starting with an initial voiceless aspirated consonant [k^h] showed a significantly closer timing to AVO with maxAcc than minVel, whereas the other three emphasized words showed a closer timing of minVel with AVO.

Concerning the results of this pilot study, a question is why one speaker showed a closer timing of maxAcc with AVO while the other two speakers showed a closer timing with minVel. Two possibilities occur: one is that the male speaker A00 had a one cm larger head than the other male speaker A05. In that Muto and Kanazawa (1996) report significant correlation between head size and amount of mandible opening, an assumption is that a larger articulator would move more slowly, and thus possibly affect the timing of articulation with AVO. This needs to be investigated further. Another possibility is that even though only one word in each utterance was written in capital letters, speaker A00 often produced more than one word in the utterance with contrastive emphasis. Future work will involve perception tests to assess this possibility. Another question is why did the word, *CAT*, which starts with a voiceless aspirated stop, show maxAcc closest to AVO, not minVel. One possibility to be investigated is the effect of VOT on mandible movement landmarks and AVO (see e.g. work by Matsui 2017).

The preliminary findings indicate that minVel of the mandible shows the best alignment with AVO; however, speaker differences as well as perhaps voicing of initial consonants can affect the timing of the landmarks. Future work will examine the interplay between segmental and syllabic articulation, specifically how crucial articulators of syllable onset and coda interact with mandible movement landmarks for a larger number of speakers. One additional contribution of the pilot study presented here is the introduction of easily to use Praat scripts available upon request for researchers to analyze acoustic/articulatory data. Tools to analyze acoustic and articulatory organization have relevance to e.g., clinical work involving speech disorders and problems with delayed language acquisition.

5. Acknowledgements

This work was supported in part by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (C) #22520412 and (C) #25370444.

6. 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. (2023). 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.

Matsui, M. F. (2017) On the Input Information of the C/D Model for Vowel Devoicing in Japanese† Michinao F. MATSUI* On the Input Information of the C/D Model for Vowel Devoicing in Japanese. J. of Phonetic Soc. of Japan, 21.1, 127–140.

Muto, T. and Kanazawa M. (1996) The relationship between maximal jaw opening and size of skeleton: a cephalometric study J of Oral Rehabilitation, https://doi.org/10.1111/j.1365-2842.1996.tb00807.x

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. (2024). Segmental and syllabic articulations: a descriptive approach. J. Speech Language and Hearing Res., https://doi.org/10.1044/2024_JSLHR-23-00092

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 (Japan)*
[2]*Lancaster University (UK)*
maho.morimoto.jp@gmail.com, t.nagamine@lancaster.ac.uk

Abstract

The current study investigates the articulation of coronal consonants /d/ and /r/ in Japanese. Using ultrasound, we obtained midsagittal tongue images for /d/ and /r/ in three phonological contexts from one male Japanese speaker. Based on the tongue shapes, time-varying changes were analyzed quantitatively using the Principal Component Analysis (PCA). Results suggest that /d/ and /r/ may differ in terms of tongue retraction and dorsal stabilization, while also supporting previous results showing the effect of the surrounding environment. The study demonstrates that quantitative articulatory analysis combining ultrasound and PCA is a useful approach to the spatio-temporal characteristics of Japanese coronal consonants, with implications for future research.

Keywords: ultrasound, PCA, Japanese, coronal consonants, dynamic analysis

1. Introduction

The current study examines the articulatory characteristics of the liquid consonant in Japanese by comparing the spatio-temporal properties of the coronal consonants /d/ and /r/. While Japanese /r/ is canonically realized as an alveolar tap or flap [ɾ], it also shows a wide range of phonetic variation, including stop-like realizations such as [d] and [ɖ] (Arai 1999; Arai 2013; Vance 1987). In addition, the degree of similarity between /d/ and /r/ is reported to vary depending on the context (e.g., Arai 2013; Okada 1991). While some consider Japanese /r/ to be a 'weak [d]' (e.g., Kawakami 1977), others argue that Japanese /r/ is articulatorily different from /d/ in that /r/ involves a ballistic gesture (e.g., Akamatsu 1997).

Previous articulatory research seems to indicate that Japanese /r/ is not a 'weak [d]'. For example, previous studies demonstrate that Japanese /r/ shows a retracted place of articulation compared to coronal stops /t/ and /d/, using electropalatography (EPG; Kochetov 2018) and electromagnetic articulography (EMA; Morimoto 2020). Another EPG study also highlights substantial variability of the /r/ realizations across vowel contexts, while no such variability is mentioned for /d/ (Kawahara and Matsui 2017). However, the exact articulatory mechanisms underlying the similarities and differences between /d/ and /r/ are not well understood. This is especially true for the movements of the tongue dorsum, which are not well-captured using EPG or flesh-tracking methods like EMA, due to the limited amount of information available on the shape of the tongue.

The current study aims to complement the previous discussion regarding the similarity between /d/ and /r/ in Japanese. We use ultrasound tongue imaging to capture clear images of the tongue dorsum, whose behavior may differ depending on the vocalic context. Tokens of /d/ and /r/ are produced in three vowel environments to investigate the realizations of Japanese /d/ and /r/. We analyze how tongue shape changes over time, especially the tongue dorsum, which may provide insights into the articulatory differences between /d/ and /r/.

2. Methods

We report results from one 21-year-old male speaker from Tokyo. The participant produced Japanese words containing intervocalic /d/ and /r/ in three different phonological environments: /a_a/, /a_i/ and /aɴ_o/, as shown in **Table 1**. The speaker produced each token five times in random order, resulting in a total of 30 tokens of /d/ and /r/ for analysis.

Table 1: *List of words analyzed in this study.*

Consonant	Context	Word	Gloss
/d/	/a_a/	/ada/	avenge
/r/	/a_a/	/ara/	coarseness
/d/	/a_i/	/badiː/	body/buddy
/r/	/a_i/	/bariː/	Barry
/d/	/aɴ_o/	/kaɴdou/	sensation
/r/	/aɴ_o/	/kaɴro/	honeydew

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 an UltraFit headset to minimize undesirable probe movement (Spreafico, Pucher, and Matosova 2018).

Data analysis is based on acoustically-delimited intervals. We first 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 were extracted at 11 equidistant time points during the target intervals of the 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, Lawson, et al. 2011).

To identify the primary variation in midsagittal tongue movement in /d/ and /r/, we conducted a principal component analysis (PCA) using scripts publicly available from Nance and Kirkham (2022). PCA was run based on the *z*-normalized *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/.

3. Results

Figure 1 shows time-varying changes in the midsagittal tongue shapes for /d/ (left) and /r/ (right) across the three vowel environments during the consonantal intervals. While the tongue dorsum movement seems similar in the /a_a/ context, we observe a slight difference in the shape of the tongue body between the two consonants. In addition, some qualitative differences can be found in the /a_i/ and /aɴ_o/ contexts. Overall, tongue dorsum movement is smaller for /r/ than for /d/ in the /a_i/ context. Minor differences can also be found around the tongue dorsum in the /aɴ_o/ context. Finally, there is a difference in tongue tip variation in the /a_a/ and /aɴ_o/ contexts (note, however, that our methodology does not allow for a clear visualization of the tongue tip).

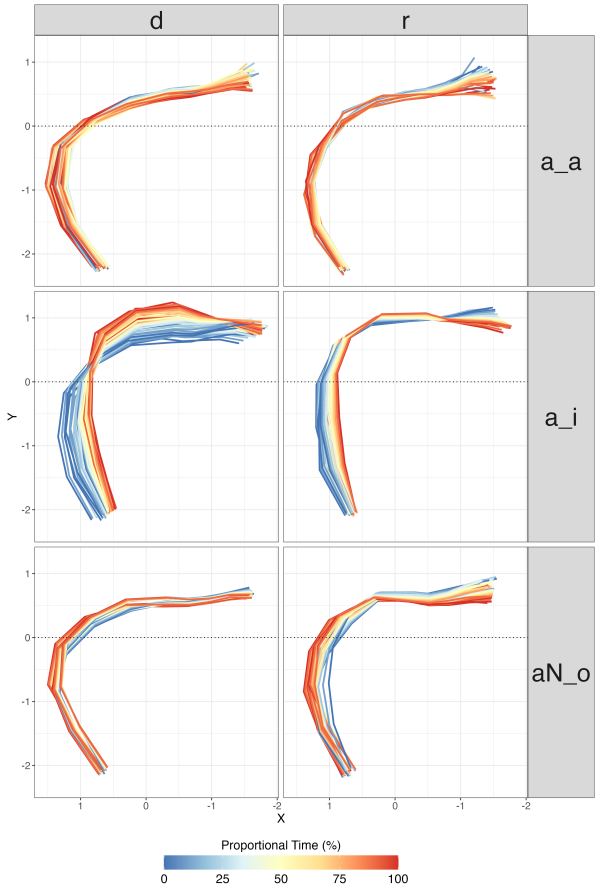


Figure 1: *Midsagittal tongue shapes in each frame during the consonant intervals in each vowel context for /d/ and /r/. Tongue tip points to the right.*

In order to explore the articulatory differences quantitatively, the results of PCA are shown in **Figure 2**. Variations explained by each principal component (PC) are superimposed on the midsagittal tongue shape, in which the mean tongue shape is represented with the bold line and the variation captured by each PC with the dashed (plus) and dotted (minus) lines by adding and subtracting a standard deviation associated with each PC from the mean tongue curve. We have found that the variation in the tongue motion of the two consonants can be described primarily in terms of two principal components, PC1 (76.85%) and PC2 (10.97%). In **Figure 2**, PC1 appears to capture the tongue retraction component at the tongue dorsum, correlated with the height of the tongue body. PC2 suggests a very subtle variation around the tongue body.

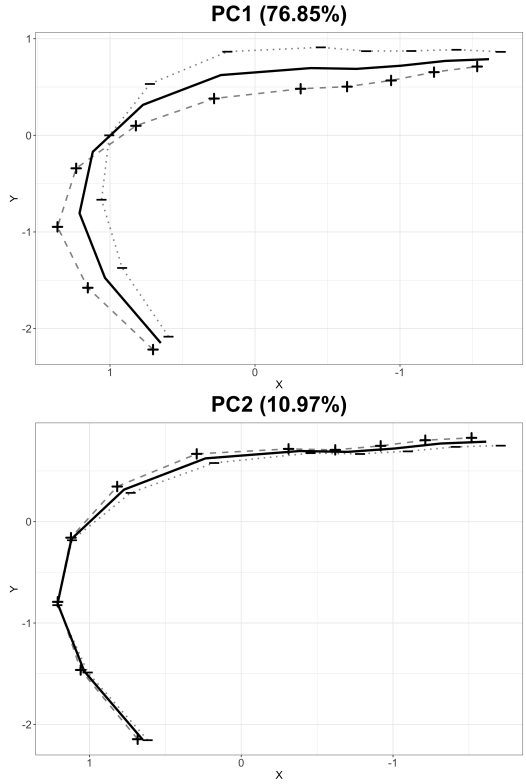


Figure 2: *Variation captured in PCs 1 and 2.*

Finally, **Figure 3** shows the changes in PC scores tracked during the consonantal intervals, allowing us to infer the articulatory movements along the PC dimensions. The consonant duration is normalized and expressed proportionally between 0% (consonantal onset) and 100% (consonantal offset). The thin lines represent PC changes of each token, with the thick lines smoothing them and the dotted lines showing the 95% confidence interval. The time-varying changes for PC1 (top three panels in **Figure 3**) 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 the /a_i/ context, 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 the /aɴ_o/ context are largely comparable with the two trajectories overlapping for the majority of consonantal intervals. Turning to PC2 (bottom three panels in **Figure 3**), the results suggest that the tongue body is slightly raised for /r/ across vocalic contexts. The difference in PC2 between /d/ and /r/ spans throughout the consonantal intervals.

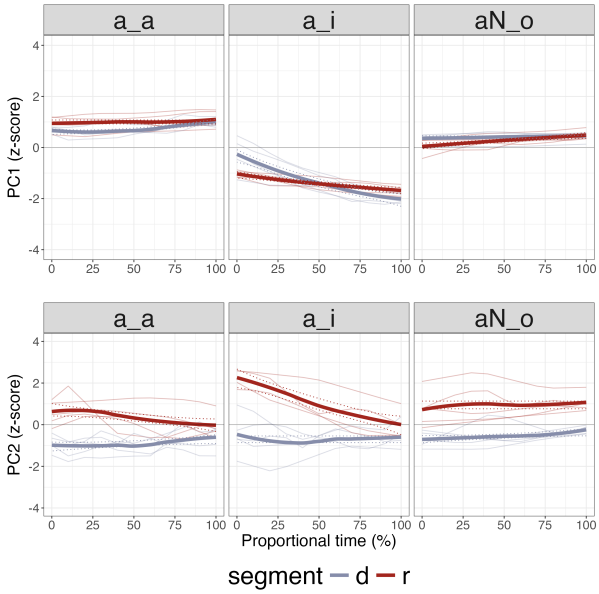


Figure 3: *Time-varying changes of each PC.*

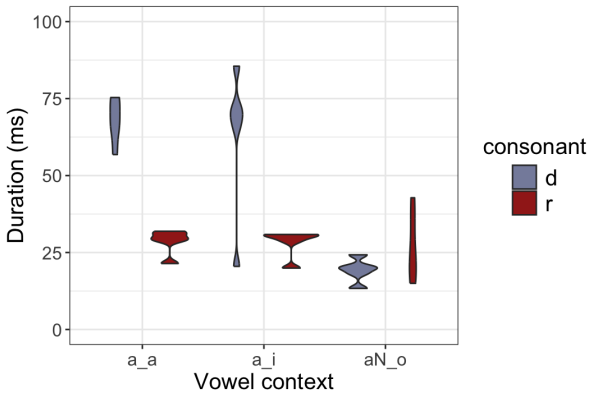


Figure 4: *Duration (ms) of the acoustic constriction of each consonant.*

4. Discussion and conclusion

The current study highlights some 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 the /a_a/ context. In addition, as seen in the midsagittal tongue shape and the dynamic changes in PC1 in **Figure 3**, the relative stability in the tongue dorsum position for /r/ in the /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 the /aɴ_o/ context seems to be in line with previous findings reporting that liquids are sometimes replaced by plosives after coda nasals in child speech, although the particular instance provided was of post-nasal /r/ replaced by /g/ (Arai 2013). This similarity may be explained by the durational differences among the phonological environments. While /d/ was generally longer than /r/ overall, we find that the duration of /d/ was quite short and thus comparable with /r/ (around 28 ms overall) in the post-nasal environment, as illustrated in **Figure 4**. We also observed this in one token of /d/ in the /a_i/ context. Spectrographic representation of this token suggests that this is an instance of the lenition of /d/, and we intend to explore the relationship between duration, lenition, and the similarity between liquids and plosives in different phonological contexts in future research. 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 could be indicative of the difference in the manner requirements for /d/ and /r/.

While it is based on a small number of tokens, the current study demonstrates that ultrasound paired with PCA allows us to investigate the articulatory mechanisms of coronal consonants. The current results seem to indicate a more stable dorsal movement for /r/ than for /d/, especially in the /a_i/ context. This could reflect dorsal stabilization as a unifying articulatory characteristic of liquid consonants (Proctor 2011), but this possibility needs to be further evaluated in future research as it is also possible that it is a result of specific manner requirements (Recasens 2016). Methodologically, note also that the vowel environment /a_i/ may exhibit a joint effect of the tongue movement and jaw displacement, making this observation inconclusive. Since the probe tracks the movement of the lower jaw, the transition of the tongue position from one vowel to another needs to be evaluated with caution (Scobbie, Wrench, and van der Linden 2008).

Nevertheless, we believe that the dynamic analysis on dorsal movement in this study is promising in identifying what articulatory mechanisms could distinguish Japanese /r/ from the coronal consonant /d/. Future research will incorporate a larger number of speakers, as the current study is based on a small number of tokens produced by a single speaker. It would also be necessary to examine the productions of /d/ and /r/ in a wider variety of contexts, as articulation of /r/ is known to be largely influenced by prosodic positions and adjacent vowels (Yamane, Howson, and Wei 2015; Maekawa 2023). Our results also suggest the need to consider the prosodic position and its effect on the duration and lenition of /d/. Furthermore, in controlling the vowel environments, we would need to take into account the dynamic jaw movement mentioned above.

To conclude, the current study provides a preliminary articulatory description of Japanese /d/ and /r/ based on ultrasound data. The results suggest that Japanese /r/ may not involve the same articulatory mechanism as /d/, highlighting key differences in the degree of tongue retraction and stabilization. Based on our observartions, we tentatively argue that Japanese /r/ is not a 'weak [d]'. The limitations of the study offer important implications for future research, which will help to achieve a better articulatory characterization of Japanese coronal consonants.

5. Acknowledgements

This study was supported by JSPS KAKENHI Grant Number JP20K21979. We thank the participant for taking part in the experiment, and Professor Takayuki Arai and the Speech Communication Lab at Sophia University for their support. We are also grateful to the anonymous reviewers of our conference abstract for pointing out, among other issues and considerations, the dynamic displacement of the probe related to jaw lowering.

6. References

Akamatsu, T. (1997). *Japanese Phonetics: Theory and Practice*. München, Newcastle: Lincom Europa.

Arai, T. (1999). "A case study of spontaneous speech in Japanese". In: *Proceedings of the 14th International Congress of Phonetic Sciences*. Ed. by J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey. San Francisco: ICPhS Archive, pp. 615–618.

— (2013). "On why Japanese /r/ sounds are difficult for children to acquire". In: *Interspeech 2013*. Lyon, France, pp. 2445–2449.

Articulate Instruments (2023). *Articulate Assistant Advanced version 221.0.0*. Articulate Instruments. Edinburgh.

Boersma, P. and D. Weenink (2022). *Praat: Doing Phonetics by Computer version 6.2.19*.

Kawahara, S. and M. F. Matsui (2017). "Some aspects of Japanese consonant articulation: A preliminary EPG study". In: *ICU Working Papers in Linguistics (ICUWPL)* 2, pp. 9–20.

Kawakami, S. (1977). *Nihongo Onsei Gaisetsu [Outline of Japanese Phonetics]*. Tokyo: Ofusha.

Kochetov, A. (2018). "Linguopalatal contact contrasts in the production of Japanese consonants: Electropalatographic data from five speakers". In: *Acoustical Science and Technology* 39.2, pp. 84–91. DOI: 10.1250/ast.39.84.

Maekawa, K. (2023). "Articulatory characteristics of the Japanese /r/: A real-time MRI study". In: *Proceedings of the 20th International Congress of Phonetic Sciences*. Ed. by R. Skarnitzl and J. Volín. Prague: Guarant International, pp. 992–996.

McAuliffe, M., M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger (2017). "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi". In: *Interspeech 2017*. ISCA, pp. 498–502. DOI: 10.21437/Interspeech.2017-1386.

Morimoto, M. (2020). "Geminated liquids in Japanese: A production study". PhD thesis. University of California Santa Cruz.

Nance, C. and S. Kirkham (2022). "Phonetic typology and articulatory constraints: The realization of secondary articulations in Scottish Gaelic rhotics". In: *Language*, pp. 419–460.

Okada, H. (1991). "Japanese". In: *Journal of the International Phonetic Association* 21.2, pp. 94–96.

Proctor, M. (2011). "Towards a gestural characterization of liquids: Evidence from Spanish and Russian". In: *Laboratory Phonology* 2.2, pp. 451–485. DOI: 10.1515/labphon.2011.017.

Recasens, D. (2016). "What is and what is not an articulatory gesture in speech production". In: *Gradus - Revista Brasileira de Fonologia de Laboratório* 1, pp. 23–42. DOI: 10.47627/gradus.v1i1.101.

Scobbie, J., E. Lawson, S. Cowen, J. Cleland, and A. Wrench (2011). "A common co-ordinate system for mid-sagittal articulatory measurement". In: *QMU CASL Working Papers* 20, pp. 1–4.

Scobbie, J., A. A. Wrench, and M. van der Linden (2008). "Head-Probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement". In: *Proceedings of the 8th International Seminar on Speech Production*. Strasbourg, France, pp. 373–376.

Spreafico, L., M. Pucher, and A. Matosova (2018). "UltraFit: A speaker-friendly headset for ultrasound recordings in speech science". In: *Interspeech 2018*. ISCA, pp. 1517–1520. DOI: 10.21437/Interspeech.2018-995.

Vance, T. J. (1987). *An Introduction to Japanese Phonology*. State University of New York Press.

Yamane, N., P. Howson, and P.-C. G. Wei (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.

`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

Abstract

This paper introduces `praatpicture`*, a library in R for making figures in the style of Praat Picture, showing one or more sound signals and a range of possible derived signals with time-aligned annotations. The library provides easy out-of-the-box solutions but also a high degree of flexibility, in many cases giving users straightforward access to changing graphical parameters that are either unavailable or relatively inaccessible in Praat. Derived signals (such as spectrograms or pitch tracks) can either be calculated on the fly using R or imported from Praat; annotations can be made interactively in R or imported from Praat. Options are available for embedding audio in figures, animating figures, and plotting annotated data directly from an EMU-SDMS database. This provides an opportunity for phoneticians who use R extensively to keep more of their workflow in a general purpose software environment.*

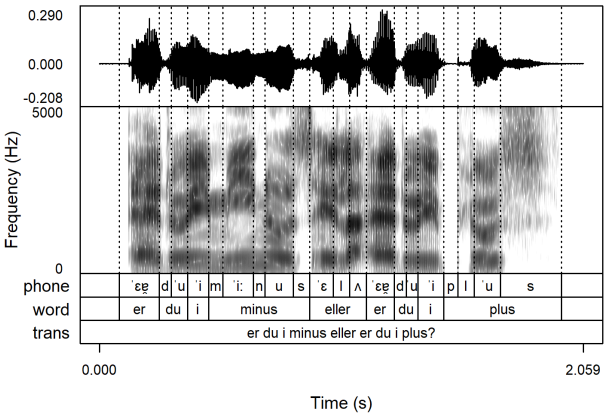
Keywords: data visualization, R, Praat

1. 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 or with scripts written in Praat's specialized custom scripting language. The GUI is highly flexible, but complicated figures have to be built incrementally, often in many steps, and it can be difficult to align figure components exactly as desired relative to each other. This can to some extent be circumvented with scripts or plug-ins, but these may not be easily accessible to the majority of potential users due to the lack of a central repository of Praat resources.

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). Due to both the lack of an out-of-the-box solution for aligning signals, derived signals, and annotations in Praat, and the widespread use of R among phoneticians, there is a need for a flexible plotting utility that aligns signals and 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 which by default resemble their counterparts in Praat as much as possible. The library also capitalizes on the possibilities available in R to provide a very high degree of graphical flexibility.



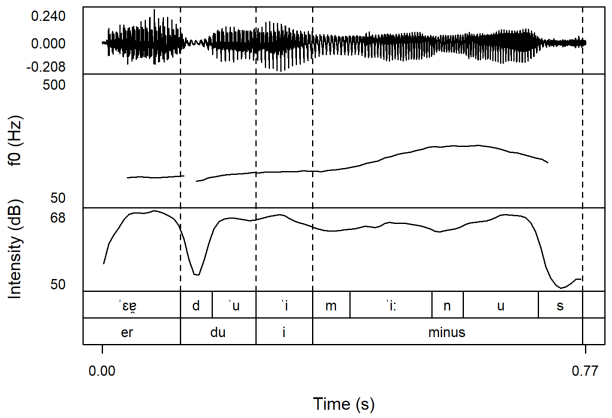

```
praatpicture(sound='ex.wav')
```

Figure 1: Simple figure generated with `praatpicture`; the grey box underneath the figure shows the code used to generate the figure.

`praatpicture` is named in tribute of Praat's plotting GUI, but it does not rely on Praat's signal processing tools and does not require a Praat installation. As discussed below, derived signals calculated in Praat *can* be used by `praatpicture`, but the library can also calculate these signals on the fly using signal processing tools that are already available in R. The library 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.

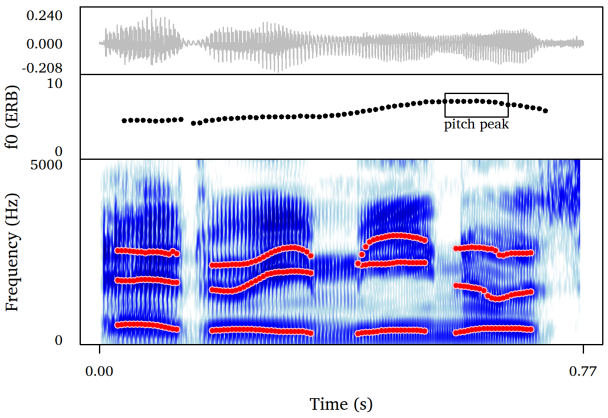
Version 1.0.0 of `praatpicture` is currently available through the central repository of R libraries, CRAN (Puggaard-Rode 2024). `praatpicture` is being continuously developed and updated.[1] A manual showing how to use all functions and parameter settings available in the package can be found at `https://rpuggaardrode.github.io/praatpicture_manual`.

[1] The plots for this paper are in fact made using version 1.1.0 of the library, which as of this writing is not yet available on CRAN, but can be downloaded from GitHub. The version update will only affect formant coloring in **Figure 3**.




```
praatpicture('ex.wav', start=0.13, end=0.9,
frames=c('sound', 'pitch', 'intensity',
'TextGrid'), proportion=c(20,40,25,15),
tg_tiers=c('phone', 'word'),
tg_tierNames=FALSE, tg_focusTier='word',
tg_focusTierLineType='dashed',
pitch_axisLabel='f0 (Hz)')
```

Figure 2: Figure showcasing some of the graphical options in `praatpicture` and the code used to generate it.




```
praatpicture('ex.wav', start=0.13, end=0.9,
frames=c('sound', 'pitch', 'spectrogram'),
proportion=c(20,25,55), wave_color='grey',
pitch_plotType='speckle' , pitch_scale='erb',
pitch_axisLabel='f0 (ERB)', spec_colors=c('white',
'lightblue', 'blue', 'darkblue'),
formant_plotOnSpec=TRUE, formant_dynamicRange=4,
formant_maxN=4, formant_color=c('red', 'pink'),
draw_rectangle=c('pitch', 0.55, 5, 0.65, 8),
annotate=c('pitch', 0.6, 4, 'pitch peak'),
family='Charis SIL')
```

Figure 3: Figure showcasing some of the graphical options in `praatpicture` and the code used to generate it.

2. Usage and options

2.1. Basic usage

The core function of the library is `praatpicture()`, which only takes one obligatory argument, `sound`, which is the name of a sound file with the `.wav` extension. Calling `praatpicture()` with just one argument will produce a very common figure format: one or more waveforms, a spectrogram, and annotations with dotted vertical lines in the various figure components indicating the locations of annotation boundaries, assuming that annotations are available (see **Figure 1**).[2] If a file with the `.TextGrid` extension which shares the same basename as the `.wav` file is available in the same directory as the `.wav` file, this will be read into R used for plotting the annotation frame. It is also possible to create multi-tiered time-aligned annotations interactively in R using the `make_TextGrid()` function.

Axis label defaults follow Praat defaults, and as in Praat, only the lowest and highest values are shown along the axes. Unlike in Praat, annotation tiers are named in the figure. Following Praat defaults, the default spectrogram is grey-scale, showing a frequency range between 0–5,000 Hz. It is generated on the basis of 1,000 spectra, each of which are calculated by applying the fast Fourier transform (FFT) to a 5 ms Gaussian window. Coloring of the spectrogram is based on a dynamic range of 50 dB, which is somewhat lower than the current Praat defaults of 70 dB. All these parameters, and many other settings, can be controlled by the user.

In the following sections, I briefly cover some of the options available to users of the package, and visualize some of these with accompanying code. Documentation for all functions and options is available using the `help()` function in R.

[2]This is a Danish sentence. Annotations were generated and force-aligned using Autophon DanFA 3.0 (Young and McGarrah 2023).

2.2. Graphical options

By default, an entire sound file is plotted, but users can specify exactly which part of a sound file to plot. The user also controls the relative size of individual plot components. Options are available to control the appearance and labels for the axes of different plot components. Users also control *which* plot components to include; **Figure 1** shows a waveform, spectrogram, and annotations, but other options are pitch tracks, formant tracks, and intensity tracks, each of which can be shown separately or overlaid on a spectrogram.

Users control which annotation tiers are plotted, and have a great deal of control over the appearance of both the annotations themselves and the vertical annotation lines shown throughout figure components. It is possible, for example, to show these lines for some annotation tiers and not others, and to vary the color and line type depending on which annotation tier they are based on. Some of Praat's special typesetting shortcuts for converting parts of annotations to boldface, italics, subscripted, small capitals etc. are also optionally available, allowing users to render annotations made in Praat according to these conventions.

Figure 2 illustrates some of these options; this figure shows a subset of the sound file visualized in **Figure 1**. The waveform is smaller, and pitch and intensity tracks are shown instead of the spectrogram. The annotation frame is also smaller, and contains only two out of three annotation tiers, without tier names printed along the side. Vertical lines shown throughout all plot components indicating annotation boundaries are based on the second annotation tier rather than the first, and the lines are 'dashed' rather than 'dotted'.

As in Praat, pitch tracks and formant tracks can be either 'drawn' or 'speckled'. Axis limits can be freely controlled for all plot components, and for pitch tracks, a range of scales are

available, *viz.* raw frequency, log frequency, semitones, ERB, and mel. The dynamic range for formant tracks and the spectrogram can also be controlled.

Users can freely adjust the colors of different plot components, and the color range to be used for spectrograms is fully customizable. There are also several options available for highlighting portions of a figure, including drawing rectangles and arrows, adding straight lines, and adding text labels. Furthermore, since the plots are made in base R, users can also access all of the base R plotting functionality, including controlling background colors, line widths, font sizes, font types, etc.

Figure 3 illustrates some of these options. This figure shows a small waveform plotted in grey. Below, there is a 'speckled' pitch track shown in the ERB scale, with a rectangle and an annotation directly on the plot component indicating the approximate location of the pitch peak. Below this, a spectrogram is plotted in hues of blue. A formant track is overlaid on the spectrogram in red colors, with a very low dynamic range of 4 dB to ensure that formants are only plotted for vowels, and with just three formants estimated (see below). The text on this figure is typeset using the Charis SIL font.

2.3. Signal processing

In addition to having a great deal of control over the look of a figure, users also have a great deal of control over the signal processing underlying the derived signals. Whenever possible, the default settings for the signal processing parameters are identical to those in Praat. Users can freely control the window length of each spectrum that makes up the spectrogram, making it possible to plot both narrowband and broadband spectrograms. It is also possible to control the number of spectra that make up a spectrogram, and the window function applied to these.

For pitch tracking, it is possible to control the measurement interval as well as floor and ceiling values. For formant tracking, it is possible to control the measurement interval, the number of formants to be estimated, and the duration of the analysis window. In both cases, but particularly in the case of pitch tracking, Praat allows quite a bit more control over the signal processing implementation than `praatpicture`. For intensity tracks, users can control the measurement interval and minimum pitch frequency. As in Praat, these derived signals are all estimated by applying Gaussian-like window functions to the analysis windows.

In addition to estimating these derived signals directly in R, it is also possible to import derived signals from Praat or any other software. I return to this option in **Section 3** below, where I also discuss differences in signal processing in more detail.

2.4. Miscellaneous functionality

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. Instead of taking the obligatory argument `sound`, `emupicture()` takes the obligatory arguments `db` (an EMU database loaded into R) and `bundle`, which is EMU-SDMS terminology for a sound file and associated annotation and signal files. Otherwise, `emupicture()` takes the exact same arguments as `praatpicture()`.

Another function `talking_praatpicture()` creates a simple single-frame video file showing a `praatpicture`-style image with embedded sound; these can either be shown in the `Viewer` pane for users of RStudio, or can be saved as MP4

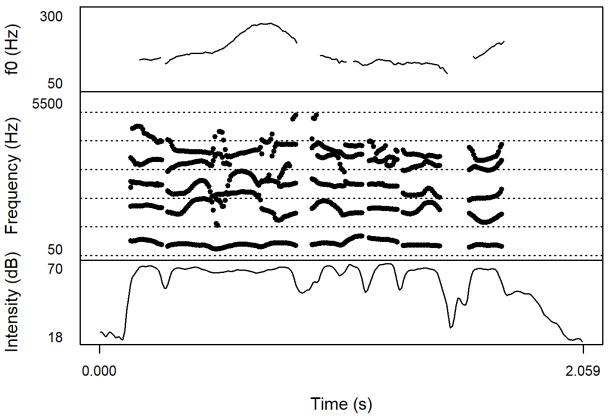


Figure 4: Three derived signals calculated using `wrassp` with the default settings in `praatpicture`.

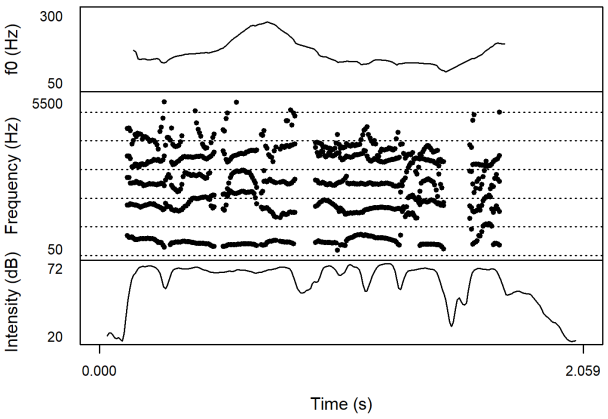


Figure 5: Three derived signals calculated using Praat with the default settings from `praatpicture`.

files. This provides a very easy way of including sound with accompanying visualization in teaching materials or conference presentations.

Finally, the function `praatanimation()` allows users to easily create `praatpicture()`-based animations. The arguments in `praatanimation()` are mostly identical to those in `praatpicture()`, but allow for e.g. two values instead of one to be passed to arguments such as `start` and `end`; this will create an animation between those two values, i.e. a video that moves through the sound file. These animations can be made on the basis of any `praatpicture()` argument that takes a continuous numeric variable, including e.g. frequency ranges, dynamic ratios, and window lengths.

3. Implementation

The spectrograms plotted by `praatpicture()` are generated in R using the `phonTools` package (Barreda 2023). Spectrograms are plotted using raster graphics, which makes it significantly faster than other methods for plotting spectrograms available in R.

Other derived signals are generated using the `wrassp` library in R (Winkelmann, Bombien, et al. 2023). Pitch is calculated using the method proposed by Schäfer-Vincent (1983),

implemented in the `ksvF0()` function. Formants are calculated using Willems' (1987) implementation of the split-Levinson algorithm as implemented in the `forest()` function. Intensity is calculated by taking the short-term root-mean-squared amplitude of the raw signal, as implemented in the `rmsana()` function. None of these functions are called using their default values in `wrassp`, but are instead called using the default values in Praat, including the use of Gaussian-like window functions (Kaiser-20 windows) instead of Hamming windows.[3] However, results will differ somewhat from Praat, since the algorithms differ. Pitch tracking in Praat is done by implementing Boersma's (1993) autocorrelation method, and formants are tracked in Praat using Burg's (1975) algorithm. The split-Levinson algorithm for formant estimation is also available in Praat, although the Praat manual explicitly recommends using the Burg algorithm instead due to its greater accuracy. Intensity should theoretically be very similar in `wrassp` and Praat, but may still differ due to slight differences in how the algorithm is implemented.

Figure 4 shows three derived signals from the sound clip also used for **Figure 1** calculated using `wrassp`, and **Figure 5** shows the same derived signals calculated using Praat with identical parameter settings wherever possible. In this particular sound file, pitch tracking fails more often when using `wrassp`; formant tracking with `wrassp` is rather less erratic than with Praat; and intensity tracking is very similar across implementations.

`praatpicture` can also plot derived signals that are calculated using Praat. If derived signals from Praat are saved to the same folder as the `.wav` file, using the same base file name and the `.PitchTier`, `.Formant`, and `.IntensityTier` extensions, respectively, then these are loaded into R and used for plotting. Alternatively, any other signal processing software can be used to calculate these signals, as long as they are stored in the Simple Signal File Format (SSFF) (Winkelmann 2017) and read into R using `wrassp`. In principle, this functionality also allows users to plot other signals, including e.g. articulatory trajectories aligned with audio, annotations, and derived signals.

Praat-based signal files and `.TextGrid` files are read into R using the `rPraat` library (Bořil and Skarnitzl 2016).

4. Conclusion

`praatpicture` provides an opportunity for phoneticians who use R 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. While R does not have the same flexibility as Praat in terms of signal processing, using base R graphics tools to produce these figures arguably has a number of advantages in terms of graphical flexilibity, and derived signals can easily be imported from Praat if the user wishes to do so. `praatpicture` currently has many of the same options as the Praat Picture GUI does in terms of producing figures with time-aligned acoustic signals and annotations, and provides basic options for annotating short sound files interactively in R, as well as functions for embedding audio in figures and producing animations.

[3] In some cases, defaults differ in `praatpicture` and Praat due to recent changes in Praat's default settings. The defaults in `praatpicture` are kept stable for backwards compatibility.

5. Acknowledgements

A huge thanks to Katie Jepson for being an early adopter of the library, for finding and pointing out several bugs in earlier versions of the library, and for suggesting many important additions. Thanks to Katie, James Kirby, Josie Riverin-Coutlée, Francesco Burroni, and Jai Peña for their help improving this paper. Thanks to audiences in Munich and Lancaster who saw demonstrations of earlier development versions of the library and provided valuable feedback. Finally, thanks to Søren Sandager Sørensen, whose voice is shown in all figures throughout this paper.

6. 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 (1993). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound". In: *Institute of Phonetic Sciences, University of Amsterdam, Proceedings* 17, pp. 97–110.

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`.

Burg, John Parker (1975). *Maximum entropy spectral analysis*. (PhD dissertation, Stanford University).

Puggaard-Rode, Rasmus (2024). "`praatpicture`. Praat Picture style plots of acoustic data". (Version 1.0.0). URL: `https://CRAN.R-project.org/package=praatpicture`.

R Core Team (2023). "R. A language and environment for statistical computing". (Version 4.3.2). URL: `https://R-project.org`.

Schäfer-Vincent, Kurt (1983). "Pitch period detection and chaining. Method and evaluation". In: *Phonetica* 40.3, pp. 177–202. DOI: `10.1159/000261691`.

Willems, Lei F. (1987). "Robust formant analysis for speech synthesis applications". In: *Proceedings of the European Conference on Speech Technology* 4, pp. 1250–1253. DOI: `10.21437/ECST.1987-18`.

Winkelmann, Raphael (2017). *The EMU-SDMS*. (PhD Dissertation, Ludwig-Maximilians-Universität Munich).

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.

Young, Nathan J. and Michael McGarrah (2023). "Forced alignment for Nordic languages. Rapidly constructing a high-quality prototype". In: *Nordic Journal of Linguistics* 46.1, pp. 105–131. DOI: `10.1017/S033258652100024X`.

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

Abstract

This study explores the articulatory underpinnings of the alleged contrast between labialized velar stops /k^{w}ai/ and onglides /kui/ in Hong Kong Cantonese, as represented in the Jyutping transliteration system. We analyzed the coordination patterns of onsets and vocoids in these syllable types as well as in canonical CV syllables. Our findings indicate that the onsets and vocoids in labialized velar stops, onglides, and CV syllables are timed synchronously, showing no variation in stiffness. These results challenge the assumed distinctions between labialized velar stops and onglides, confirming the presence of CV synchrony in Hong Kong Cantonese.

Keywords: Gestural coordination, Prenuclear glides, labialized velar stops, Hong Kong Cantonese, EMA

1. Introduction

The status of prenuclear glides remains a contentious issue in Chinese phonetics and phonology. Descriptively, the maximal syllable structure in most Sinitic languages can be represented as CGVX, where C represents a consonant, G a glide, V a vowel, and X can be either a stop consonant or another glide (Duanmu, 2007). However, the prenuclear glide, referred to as *Jièyīn* 'the medial sound' in Chinese historical phonology, presents a significant anomaly within this framework. The primary challenge lies in the ambiguity surrounding the analysis of CG sequences. These sequences can be interpreted as complex segments (e.g., labialized consonants), onglides, or independent "medial sounds." Unfortunately, no empirical evidence definitively supports any of these analyses. As a matter of fact, Myers's (2015) comprehensive survey highlights the perplexing diversity of findings from various approaches, including phonotactics, acoustic measurements, rhyming patterns, language games, syllable manipulation experiments, acceptability judgment tasks, speech errors, and first language acquisition data. It is fair to say that no universally applicable analysis exists that can reconcile all these mutually contradictory results (see also Van der Weijer & Zhang, 2008).

By contrast, Cantonese seems to offer a clearer picture. Most syllables conform to a simpler CVX template (where C = consonant, V = vowel, X = stop consonant or glide). The only "exceptions" to this generalization are syllables beginning with *kw-* and *gw-*. For example, *kwai1* 'rules' or *gwan1* 'military' deviate from the CVX template. Such "anomalies" have led scholars, including Lin (1990), to propose that *kw-* and *gw-* should be analyzed as single units — or "co-articulated onset," as termed by Bauer & Benedict (2011) — specifically as labialized velar stops (i.e., /k^{w}/ and /g^{w}/). This analysis is, to our knowledge, primarily based on the very assumption that the maximal syllable in Cantonese is supposed to be CVX.

Nevertheless, the widely accepted analysis is not without its limitations. While labialized velar stops neatly conform to the CVX template, this analysis has faced criticism for relying predominantly on phonological considerations rather than substantial empirical evidence. More critically, this treatment does not adequately address syllables that exhibit onglides, such as *bui1* 'cup' and *kui1* 'hinoki cypress.' An astute reader might question why *kui* 'hinoki cypress' is not transcribed as *kwi* in this instance. In fact, our transcription follows the *Jyutping* system, or the Linguistic Society of Hong Kong Cantonese Romanization Scheme. Developed in 1993 by the Linguistic Society of Hong Kong, *Jyutping* provides a standardized method of Romanizing Cantonese. Within the *Jyutping* system, *kwai1* 'rules' and *kui1* 'hinoki cypress' are treated as distinct structures: the former is represented as a consonant-like element and the latter is regarded as a vowel-like component, forming part of a diphthong. Once again, this distinction is presumably based on the CVX template, which analyzes the *kw-* sequences as a labialized velar stop.

Building upon the discussion above, the primary research goal of this study is to quantitatively confirm whether the distinction between labialized stops and onglides can be reliably established. To this end, we adopt the approach of Shaw et al. (2021), which suggests that complex segments (in this case, labialized velar stops) and segment sequences (onglides) can be differentiated by distinct patterns of intergestural coordination, as detailed in Section 2.4. Beyond these "extrinsic" differences such as coordination, we further explore "intrinsic" differences between vowels and glides (see Burgdorf & Tilsen (2021) for a recent attempt to quantify such differences in American English). Specifically, we focus on the metric of stiffness as it has been hypothesized that "[t]he consonantal gestures typically exhibit greater degrees of constriction and shorter time constants (higher stiffness) compared to vocalic gestures" (Browman & Goldstein, 1992: 30).

In sum, the present study aims to provide empirical evidence to either support or challenge the purported phonological contrast between these two classes of Cantonese sounds: the labialized velar stop, as exemplified in *kwai1* 'rules' versus the onglide, as in *kui1* 'hinoki cypress' or *bui1* 'cup.'

2. Methods

2.1. Participants

This study involved eight (8) native Cantonese speakers from Hong Kong, three of whom were female. All participants were in their twenties at the time of data collection. These individuals were born and raised in Hong Kong and reported no history of hearing or reading impairments. Participation was voluntary and participants received compensation for their involvement.

2.2. Materials

The target syllables were the initial syllables in a disyllabic word (e.g., *bui1 dip6* 'cup (and) dish'). Three categories of the target items were examined: (i) **C^{G}**: labialized velar stops {*kwai*, *gwong, gwai*, and *gwaai*}, (ii) **CG**: syllables with diphthongs *ui* or *iu*: {*fui*, *pui*, *bui*, *kui*; *piu*, *biu*, *tiu*, *diu*, *giu*, *siu*, and *ziu*}, and (iii) **CV**: syllables with monophthongs {*bun*, *gong*, *bi*, *gu, gun, bing, ding,* and *ging*}. All target items carried level tones and

were embedded in the carrier phrase: *gaa __ bei keoi* [kaa __ pei kʰøi], meaning 'add __ for him/her.' In total, 1,840 tokens (= 23 items × 10 repetition × 8 speakers) were analyzed and reported in this study.

2.3. Data Acquisition

Kinematic data were collected at the phonetics laboratory of the National Tsing Hua University using a Carstens AG 501 electromagnetic articulograph (EMA). This data collection adheres to the experimental protocols outlined in Rebernik et al. (2021). Additionally, EMA sensors were affixed to the speaker's tongue in the "Southern Cross" configuration, as described by Ying et al. (2021). This specific sensor placement was chosen because the experiment also involved a study of lateralization in Hong Kong Cantonese.

2.4. Analysis

The EMA data was processed using MView (Tiede 2005) and custom MATLAB scripts. Articulatory gestures were identified using the *findgest* function in MView. The following trajectories were used for gestural analysis: for onset consonants, we used the lip aperture (LA) trajectory to identify labial gestures, the tongue tip (TTz, where 'z' indicates vertical movement) trajectory for alveolar gestures, and the tongue dorsum (TDz) trajectory for dorsal gestures. For vocoids, the tongue body (TBz) trajectory was employed to pinpoint palatal gestures, while the tongue dorsum (TDz) and upper lip (ULx, where 'x' indicates longitudinal movement) trajectories were used to identify /u/ and /w/. Additionally, the labial onset of *bei* 'for' in the carrier phrase served as the anchor for normalization.

Interval normalization is depicted schematically in **Figure 1**. To compute the duration of G1, we subtracted the time of G1 onset from G1 offset. Similarly, the onset lags between G1 and G2 (hereafter referred to as "onset-to-onset lags") were determined by calculating the difference between G1 onset and G2 onset. These calculated intervals were then normalized by dividing each by the total interval from G1 onset to Anchor onset.

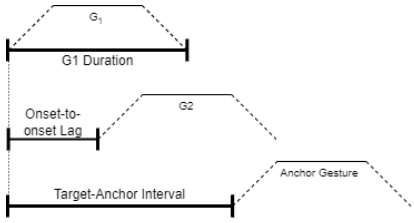


Figure 1: *Interval normalization.*

Following Shaw et al.'s (2021) methodology, we hypothesize that the relationship between G1 duration and onset lags can be indicative of distinct segmental compositions. Specifically, if G1 duration and onset lags are found to be independent, exhibiting weak or no correlation, the segmental composition will be analyzed as a complex segment, as shown in **Figure 2a**. Conversely, if a strong correlation is observed between G1 duration and onset lags, suggesting covariance, the analysis will classify the segmental composition as a sequence of segments (see **Figure 2b**).

In this study, stiffness is operationalized as the amplitude-normalized peak velocity of the articulatory trajectories (Roon et al., 2021). To compute stiffness values (k'), the peak velocity ($v^\wedge$) of each articulatory movement is divided by its amplitude (A), as in (1). For the purposes of this study, the stiffness of /u/ and /w/ was quantified using the TDx trajectories and the stiffness of the vowel /i/ was assessed using the TBz trajectory.

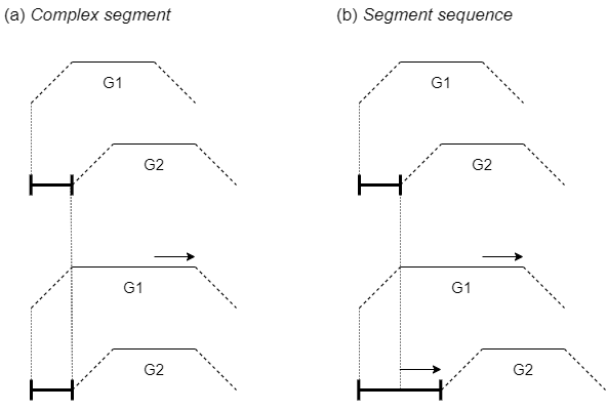


Figure 2: *Different patterns of coordination between (a) complex segments and (b) segment sequences: (a) Complex Segment: Shows stable onset lags (represented by a bold line) regardless of elongation in G1 duration; (b) Segment Sequence: Depicts onset lags increasing in proportion to the elongation of G1 duration.*

Amplitude-normalized peak velocity: $k' = v^\wedge / A$ (1)

3. Results

By employing least-squares linear regression and correlation tests, **Table 1** summarizes the coefficients of determination (R^2) across the three categories (C^G, CG and CV) under investigation. Overall, the analysis reveals no significant differences in articulatory timing between CV and CG ($R^2 \approx 0$). Conversely, the correlation analysis for C^G shows a robust correlation when measured by the tongue dorsum trajectory (TDx), yet a weak correlation is observed with the upper lip trajectory (ULx: lip protrusion). The subsequent sections further elaborate on these findings, highlighting the glaring asymmetry in the articulatory coordination of the C^G structure across the ULx and TDx dimensions.

Table 1: *Coefficients of determination (R^2)*

C: LA/TDz	C^G	CG	CV
i (TBz)	N/A	0.00	0.07
w/*u* (ULx)	0.07	0.07	0.00
w/*u* (TDx)	**0.62**	0.06	0.02

3.1. The palatal gesture

Figure 3 illustrates the timing patterns within Cantonese syllables, specifically examining the interaction between the onset consonant (G1) and the following *palatal* vocoid in CV syllables (e.g., *bi*) compared to CG syllables (e.g., *biu*). Recall that C^G is not possible in these cases (e.g., *diao* in Mandarin). The plots in **Figure 3** employ least-squares linear regression lines (in red) to illustrate the relationship between the duration of the onset consonant (G1 duration) and the onset lag (See also **Figure 2**). As shown, the regression lines for both CV and CG appear substantially flat, with R^2 approximately equal to 0, suggesting no significant trend in the data. In other words, an increase in G1 duration does *not* correspond to an increase in the onset lag. Thus, the present results indicate that the onset consonant and the palatal vocoid are produced in close temporal synchrony. The lack of covariation between G1 duration and onset lag points to a complex segmental composition in both the CV and CG contexts, as far as the palatal gesture in concerned.

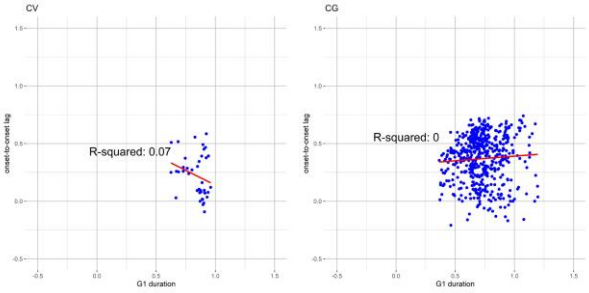


Figure 3: *Scatter plots with linear regression lines (in red) illustrating the interaction between G1 duration and onset-to-onest lags in Cantonese CV and CG configurations involving palatal gestures (bi vs. biu). The x-axis denotes G1 duration, while the y-axis indicates onset-to-onset lags.*

3.2. The labio-velar gesture

Figure 4 provides an examination of the temporal organization of labio-velar gestures in Cantonese, displaying the relation between the three putatively different categories: C^G (*kwai*), CG (*kui*), and CV (*gu*), presented sequentially from top to bottom. The present analysis looks into the trajectories of the upper lip (UL) and tongue dorsum (TD), premised on the involvement of both /w/ and /u/ regarding labio-velar gestures. The results displayed in the left-hand column of the figure, which focus on lip protrusion (ULx), reveal no significant correlation between the duration of G1 and the onset-to-onset lags ($R^2 \approx 0$) across all conditions. This lack of correlation suggests that labial gestures (or, more precisely, lip protrusion measured by ULx) are synchronized with the preceding consonant gestures, indicating consistent timing patterns across the board.

In stark contrast, the right-hand column, which assesses tongue body retraction (TDx), reveals a distinct pattern of gestural coordination. Precisely, a strong correlation ($R^2 > 0.6$) in CG configurations, such as in the syllable *kwai* 'rules', suggests asynchronous timing, indicating that the two gestures are timed sequentially {C: TDz, ^{G}: TDx}. Conversely, the analysis of CG and CV configurations, exemplified by *kui* 'hinoki cypress' and *gu* 'aunt,' shows a synchronous coordination (i.e., $R^2 \approx 0$) with the same gestures {C: TDz, G/V: TDx}.

3.3. Stiffness

Table 2 displays the average stiffness values and their corresponding standard deviations under various experimental conditions. The optimal model, derived using a generalized linear mixed-effects approach, is as follows: Condition (C^G, CG, CV) ~ Stiffness + (1 | Speaker) + (1 | Item). The analysis indicated no statistically significant differences between glides and vowels, with *p*-values exceeding 0.05 (not shown here). Comparative analysis of the trajectories involved in producing /w/ and /u/ revealed that lip protrusion (ULx) occurred at a significantly faster rate than tongue dorsum retraction (TDx). Despite these notable velocity differences, the relatively high Akaike Information Criterion (AIC = 1655.60) suggests potential overfitting of the model. Consequently, there appears to be no intrinsic difference in stiffness between /u/ and /w/ based on this study.

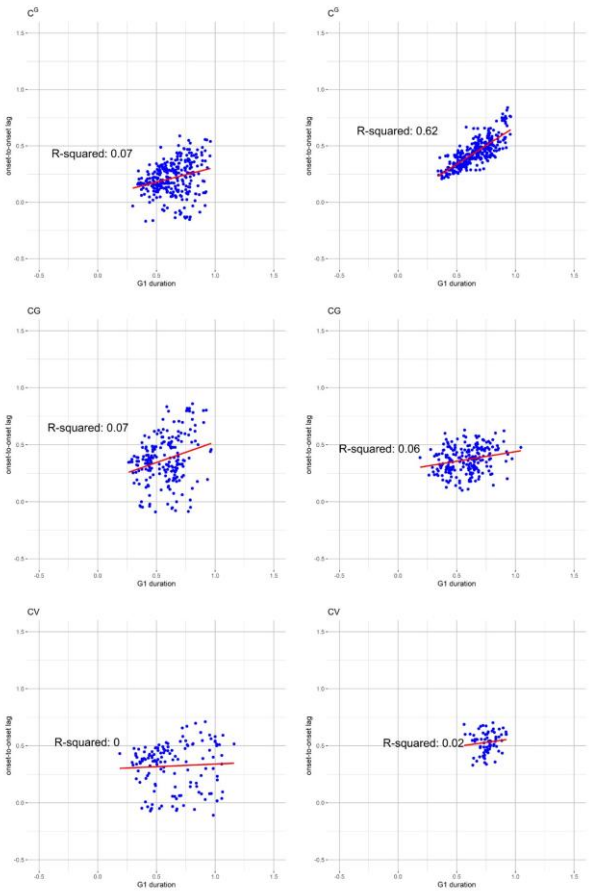


Figure 4: *Scatter plots with linear regression lines (in red) illustrating the interaction between G1 duration and onset-to-onset lags in Cantonese C^G, CG and CV configurations involving labio-velar gestures. The x-axis represents G1 (=TDz) duration, and the y-axis shows onset-to-onset lags. In the plots, ULx is used as G2 in the left-hand column, while TDx is applied as G2 in the right-hand column.*

Table 2: *Stiffness*

	Trajectory	Mean	SD
C^G (K^w)	ULx	31.41	17.41
	TDx	24.44	9.17
CG (-ui)	ULx	30.12	16.77
	TDx	20.37	10.42
CV (-u)	ULx	29.49	17.56
	TDx	20.79	11.26
CG (-iu)	TBz	21.40	11.08
CV (-i)	TBz	14.01	5.24

3.4. A note on the *-ing* rimes

Figure 5 displays the results for the target syllables with the *-ing* rimes — specifically, *bing*, *ding*, and *ging* — which were initially intended to be used to investigate CV coordination. Contrary to the *bi* syllables, these findings reveal a slightly stronger correlation ($R^2 \approx 0.3$), suggesting that the timing relations between CV and CVŋ might differ to certain extent.

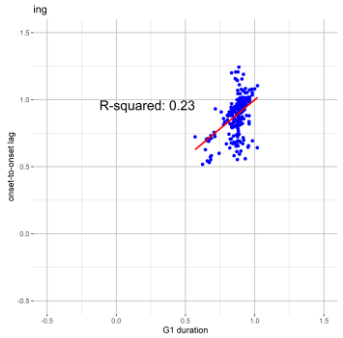


Figure 5: *Scatter plots with linear regression lines (in red) illustrating the interaction between G1 duration and onset-to-onest lags in Cantonese CVN configurations involving palatal gestures (bing, ding, and ging). The x-axis denotes G1 duration, while the y-axis indicates onset-to-onset lags.*

4. Discussion

The current results suggest that the purported distinctions among the Cantonese categories C^G, CG, and CV cannot be differentiated based on articulatory timing. Additionally, stiffness or amplitude-normalized peak velocity does not influence the vowel/glide distinction.

Our results suggest that the onsets and their respective vocoids are timed synchronously in Hong Kong Cantonese. Specifically, no significant correlations were observed between the duration of G1 and the onset-to-onset lags (see **Figure 2**). These findings challenge the prevailing assumption that CV synchrony is absent in tone languages, an assumption supported by the presence of a positive CV lag in Mandarin CV syllables (see Shaw (2022) for a recent review). In contrast, our results support the findings of Liu et al. (2022) on Mandarin, which also demonstrated CV synchrony using a minimal triplet paradigm. Future research should explore whether these conflicting findings arise from differences in experimental paradigms, as Svenssen Lundmark et al. (2021) have suggested.

This study is significant in that it involved estimating the onsets and offsets of gestural control using movements from relatively dependent articulators, such as *tiu* {C: TTz; G: TBz}. This methodology has proven effective, allowing for reliable assessment of timing relations even under these conditions. More crucially, the sole instances of "aberrant" coordination patterns displaying asynchrony, namely C^G (i.e., *kwai* and *gwai*), occur within the same sensor dimensions: {C: TDz; ^{G}: TDx}. It is hypothesized that the pharyngeal vowel /a/ in *kwai* and *gwai* could induce a more pronounced tongue root retraction. In contrast, similar syllables such as *kui* and *gu* do not necessarily involve tongue root retraction hence CV/CG synchrony. Finally, it should also be noted that the high front vowel /i/ undergoes lowering in the context of a velar nasal coda in Hong Kong Cantonese (refer to Bauer and Benedict (2011) and references cited therein). We suspect that this phonotactic constraint might also have implications for the results pertaining to the *-ing* rimes reported in in Section 3.4. In sum, the findings from the Cantonese data imply that CV or CG coordination is potentially influenced by adjacent gestures, which necessitates further investigation in the future.

5. Conclusion

The principal findings of this study challenge the distinctions delineated in *Jyutping* transliteration. As demonstrated in the preceding analysis, it is evident that the onsets and the vocoids in C^G, CG, and CV uniformly exhibit full synchrony in their gestural coordination. In conclusion, these categories do not appear to be distinguishable based on their articulatory characteristics.

6. References

Bauer, R. S., & Benedict, P. K. (2011). *Modern Cantonese phonology* (Vol. 102). Walter de Gruyter.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155-180.

Burgdorf, D. C., & Tilsen, S. (2021). Temporal differences between high vowels and glides are more robust than spatial differences. *Journal of Phonetics*, *88*, 101073.

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

Lin, Y. H. (1990). Prenuclear glides in Chinese. Mid-America Linguistics Conference.

Liu, Z., Xu, Y., & Hsieh, F. F. (2022). Coarticulation as synchronised CV co-onset–Parallel evidence from articulation and acoustics. *Journal of Phonetics*, *90*, 101116.

Myers, J. (2015). Stuck in the middle: Mandarin medials in articulation, parsing, and association. In Y. E. Hsiao & L.-H. Wee (Eds.) *Capturing phonological shades within and across languages* (pp. 101-119). Cambridge, UK: Cambridge Scholars Publishing.

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology*, *12*(1), 6.

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.

Shaw, J. A. (2022). Micro-prosody. *Language and Linguistics Compass*, *16*(2), e12449.

Svensson Lundmark, M., Frid, J., Ambrazaitis, G., & Schötz, S. (2021). Word-initial consonant–vowel coordination in a lexical pitch-accent language. *Phonetica*, *78*(5-6), 515-569.

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

Van de Weijer, J., & Zhang, J. (2008). An X-bar approach to the syllable structure of Mandarin. *Lingua*, *118*(9), 1416-1428.

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.

Production Allophones of North American English Liquids

Mark Tiede[1], Suzanne Boyce[2], Michael Stern[3], Teja Rebernik[4], Martijn Wieling[4]

[1]*Dept. of Psychiatry, Yale University & Haskins Laboratories, New Haven, USA*
[2]*Dept. of Communication Sciences and Disorders, University of Cincinnati, Cincinnati, USA*
[3]*Dept. of Linguistics, Yale University, New Haven, USA*
[4] *Dept. of Information Science, University of Groningen, Groningen, Netherlands*

mark.tiede@yale.edu, boycese@ucmail.uc.edu, michael.stern@yale.edu, t.rebernik@rug.nl, m.b.wieling@rug.nl

Abstract

The syllabic liquids [ɚ] (as in "purr") and [əɫ] (as in "pull") have well-defined acoustic targets but are produced with a wide range of heterogenous tongue postures. This work surveys midsagittal tongue shapes from a large (N=78) number of speakers producing these sounds, to illustrate their variety, and to determine systematically how this variety can be quantified. In particular we propose that a categorization based on just two parameters—degree of tongue dorsum convexity and tip orientation—is sufficient to classify observed shapes, and superior to defining *ad hoc* prototypes.

Keywords: rhotics, liquids, speech production, MRI, ultrasound

1. Introduction

The North American English (NAE) syllabic liquids /r/ [ɚ] (as in "purr") and velarized (dark) /l/ [əɫ] (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.

2. Methods

2.1. Participants

Midsagittal imaging data were collected from two non-overlapping cohorts during production of syllabic /r/ and /l/. The first group was imaged in supine posture using magnetic resonance imaging (MRI) at the University Hospital of the University of Cincinnati. The second group was imaged by stabilized ultrasound in sitting posture using the facilities of the mobile SPRAAKLAB (Wieling et al. 2023). In total 78 speakers (39F) provided the data surveyed here, ranging in age from 16 to 68 (mean 34.8, s.d. 12.6).

2.1.1. MRI

29 native NAE speakers (10F) 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. All provided informed consent and were compensated for their participation.

2.1.2. Ultrasound

To increase power, an additional 70 Dutch speakers were recorded in SPRAAKLAB producing five repetitions of (English) "purr" and "pull" with midsagittal ultrasound during the 2022 Noorderzon Festival (Groningen) using the UltraFit probe stabilizer (Spreafico et al., 2018), recorded with synchronized audio by AAA software (Articulate Instruments). The imaging frame of 720x540 pixels mapping 4.7pixels/mm was recorded at 82 frames/sec. Speakers provided informed consent but were uncompensated volunteers. Following review by two native English listeners 21 of these participants were excluded for inconsistency across repetitions or productions that did not achieve native formant targets, retaining 49 speakers (29F, 1 Other).

2.2. Analysis

2.2.1. MRI-specific

Midsagittal tongue shapes for /r/ and /l/ were obtained by fitting a thin plate spline to the lingual surface, from the top of the epiglottis to the anterior-most point of the apex. Four landmarks were identified along the distal vocal tract wall (base of the pharynx, anterior apex of the second vertebra, highest visible point of the palatal vault, and base of the alveolar ridge), and used to define a semipolar grid to 'unwrap' the tract (Figure 1). Distance functions sampled along gridlines were 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, addressing the question of how many distinct classifications of /r/ and /l/ were present in this data sample.

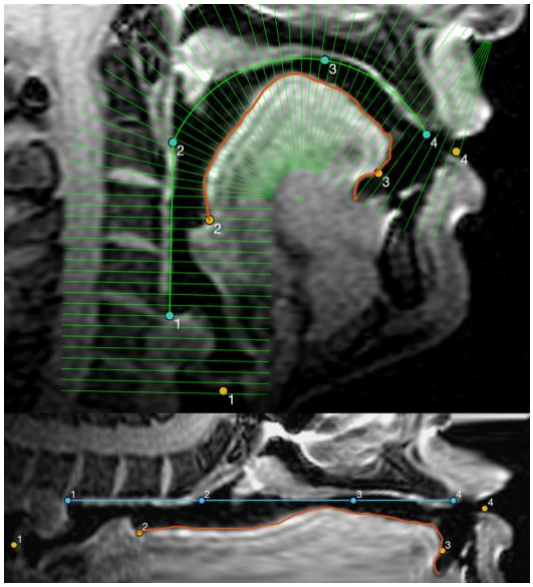


Figure 1: *Semipolar grid (above) used to sample vocal tract distance function along 'unwrapped' tract (below).*

2.2.2. *Ultrasound-specific*

Phone segmentations of the productions of "pull" and "purr" were identified using the Montreal Forced Aligner (McAuliffe et al., 2017). Tongue surface contours at the centers of these acoustically determined liquid intervals were extracted from the ultrasound video using DeepEdge (Chen et al., 2020). Three consecutive frames were averaged for each repetition, and these averages were in turn averaged across repetitions by speaker.

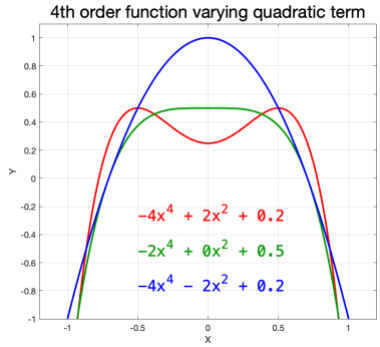


Figure 2: *Illustration of how quadratic coefficient (C2) tracks convexity: >0 concave; ~0 flat; <0 convex.*

2.2.3. *Tongue shape*

The 58 speaker tongue shapes obtained from the MRI and ultrasound cohorts for /r/ and /l/ were normalized as follows:

- resampled to an equal number of mm-based coordinates
- fitted with an ellipse enclosing 95% of all coordinates
- rotated such that the major axis of this ellipse was aligned with the horizontal coordinate axis
- 'curled-under' points at the beginning and end were trimmed (to ensure horizontal monotonicity)
- centered on the midpoint of the ellipse major axis and scaled by its length

This procedure resulted in a tongue shape y expressed as a function of x for each contour, which was parameterized by a least-squares fit to a 4th order polynomial (higher orders improved the fit but did not significantly affect the quadratic term). In addition, the rotation and scaling factors provide indices of speaker vocal tract morphology. As illustrated in Figure 2, the quadratic coefficient (C2) of this polynomial tracks the degree of convexity of the fit, and as such provides a useful characterization of tongue dorsum shape: concave shapes (bowed down/inward) have positive sign, flat shapes are close to zero, and convex shapes (bowed up/outward) have negative sign.

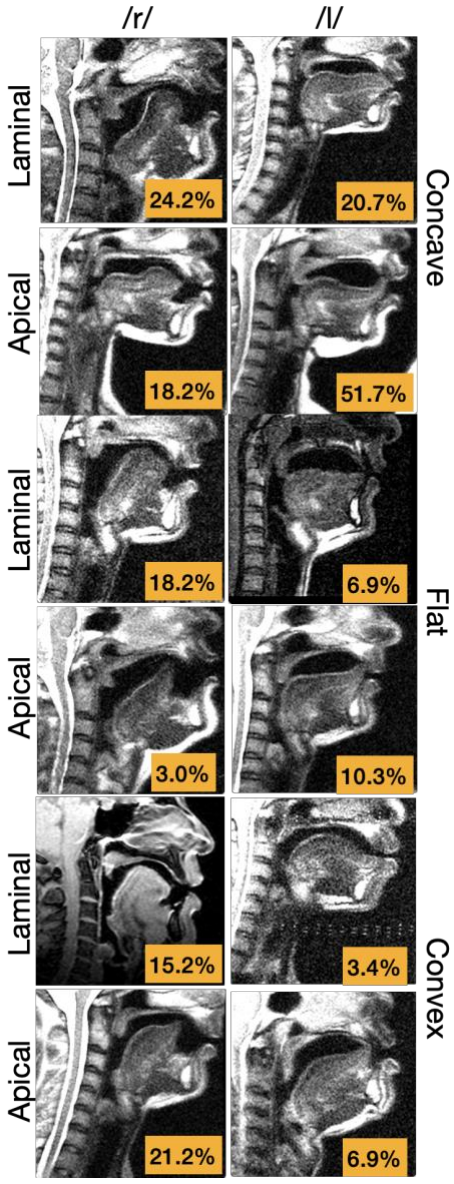


Figure 3: *Representative MRI tongue shapes for syllabic liquids showing apical and laminal tongue tip variants for concave, flat and convex tongue dorsum postures. Insets show percentage of observed speakers with that shape.*

The orientation of the tongue tip was determined with a similar parameterization. The anterior-most 30% of the original extracted tongue shapes were fitted by an enclosing ellipse, rotated, and scaled as above, though retaining non-monotonic points. The average rotation in this case is

approximately 90° CCW. Aligned in this way, the center of gravity (COG) of the polygon determined by the scaled and rotated points has negative sign in *x* for tip-up ('retroflex'/apical), and positive sign for tip-down ('bunched'/laminal) tongue shapes. Note that this secondary analysis is restricted to the 29 speakers of the MRI cohort, as the sublingual cavity and/or mandibular shadow preclude accurate imaging of the tongue tip using ultrasound.

3. Results

As a first approximation observed tongue shapes derived from MRI can be sorted into the six shapes exemplified in Figure 3. These distinguish between concave, flat and convex tongue dorsum shapes, further separated by whether the tongue tip is tilted up (apical) or down (laminal). When these shapes are characterized by Fourier decomposition of their respective distance functions as described in Section 2.2.1 above, a *k*-means classification of their associated coefficients clusters optimally into three groups by both silhouette and elbow heuristics (N=29). Principal component analysis of all speaker tongue shapes (N=78) showed independently that three components accounted for 95% of variance for both /r/ and /l/.

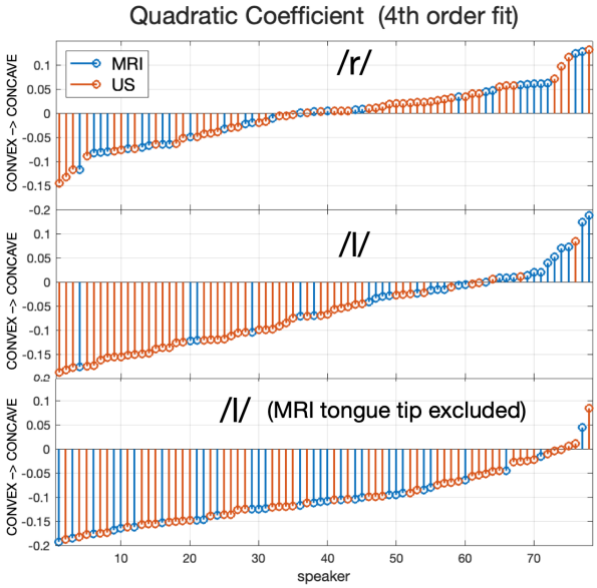


Figure 4: *Distribution of quadratic coefficients for 4th order polynomial fit to normalized tongue dorsum shapes (N=78).*

The results of fitting a 4th order polynomial to the normalized tongue dorsum data from all speakers are shown in Figure 4 (N=78). It can be seen that for /r/, shapes derived from MRI are distributed across the range about the same as those from ultrasound, confirmed by a linear model predicting C2 from data source (t(76) = 0.07 n.s.). For /l/, however, there is a strong bias towards negative (convex) shapes for the ultrasound data not seen in the MRI shapes (t(76)= -5.52 ***), which likely reflects the latter including tongue tip information not available from ultrasound. When fits are computed for /l/ with the anterior-most 30% of the MRI excluded (Fig. 4, bottom panel), this difference is no longer significant (t(76) = 1.41 n.s.), and we therefore conclude that data from both modalities can be successfully combined with this exclusion operative. Normalized tongue dorsum shapes (excluding MRI tongue tips) are shown averaged across concave, flat and convex values of C2 in Figure 5 (threshold for "flat" +/- .02).

When only complete (tongue tip included) MRI shapes are considered, there is a significant correlation between C2 values for /r/ and /l/ (r = 0.39 *). For the tongue tip, we observed that using the rhotic apical (COG<0) *vs*. laminal (COG>0) 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/ (MRI only; N = 29).

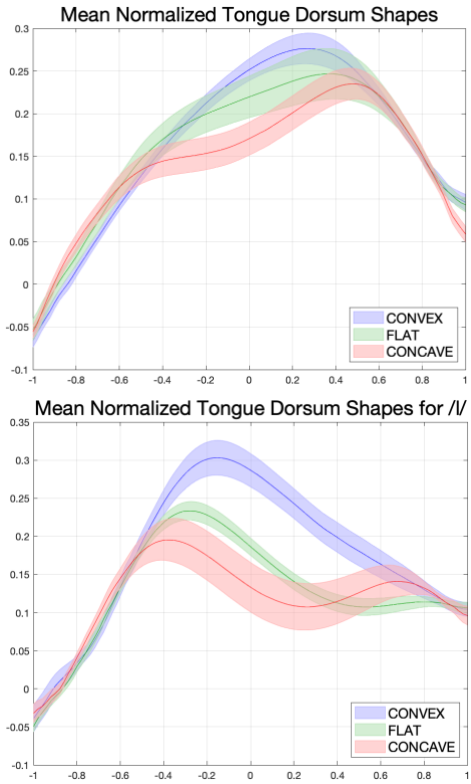


Figure 5: *Normalized tongue dorsum shapes (excluding MRI tongue tips); error bars show SEM.*

Although not directly part of the parameterization, scaling and rotation factors used to normalize tongue dorsum shapes showed an interesting gender-distinct pattern predictable from known differences in morphology (e.g. Vorperian et al., 2005): Rotation angles were consistently smaller for female speakers (t(153) = -3.23 **), likely reflecting shorter pharynx lengths relative to overall vocal tract length and thus less scope for tongue body rotation. Similarly, scale factors were also reliably larger for female speakers (t(153)= 2.71 **), likely reflecting smaller head and tongue sizes.

4. Discussion and conclusion

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, MRI 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 (lips, and two or more of the tongue within the vocal tract). 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 (/Cr/) contexts showed an apical posture during the rhotic for the former and a laminal posture for the latter (Figure 6). 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.

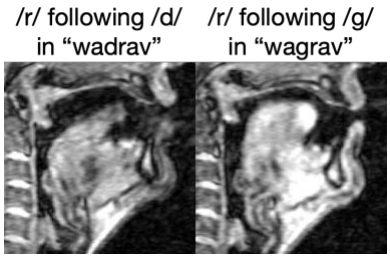


Figure 6: *Coproduced /Cr/ onset contexts from the same speaker show contrasting apical (left) vs. laminal (right) tongue postures.*

Predicting a given speaker's preferred tongue posture for liquids on the basis of their vocal tract morphology would be useful for guiding possible clinical intervention, but this remains a very challenging problem, with parasagittal shape, tongue size, muscle interdigitation and asymmetry just some of the unknown free variables affecting the observed midsagittal projection. Previous studies of midsagittal shapes of liquids have mostly followed the pioneering efforts of Delattre & Freeman (1968), who categorized the 48 shapes they observed using cineradiography into one of eight prototypes. Because our own survey found shapes that could not be readily accounted for by these prototypes, a useful step towards addressing the prediction problem is a more precise way of quantifying midsagittal shape. The two parameter approach proposed here represents an improvement over prototype classification in that it accurately separates the six basic shapes found in our survey, is arbitrarily extensible to parameterizing any midsagittal shape, and provides quantified values that can correlated with available morphological measures.

5. Acknowledgements

Work supported by NIH grants DC05250 (Boyce) and DC002717 (Whalen). Thanks to Alan Wrench for ultrasound/audio alignment, Defne Abur for assisting with data collection during Noorderzon, and especially all the participants who volunteered their time at the festival.

6. 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.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proc. Interspeech 2017* (Stockholm), 498–502.

Mielke, J., Baker, A., & Archangeli, D. (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language*, 101-140.

Spreafico, L., Pucher, M., Matosova, A. (2018). UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science. *Proc. Interspeech 2018* (Hyderabad), 1517-1520.

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.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 117(1), 338-350.

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.

Contrasting phonetic effects of morphological boundaries for vowel and consonant suffixes

Motoki Saito[1]

[1]*University of Tübingen*

motoki.saito@uni-tuebingen.de

Abstract

*Affixes (e.g., free+**s**) are sometimes found to be longer and other times shorter than pseudo-affixes (e.g., free**ze**). These opposite effects may be due to different degrees of sonority of the target segments being investigated. Vowels (or affixes containing vowels) may be easier to make acoustically more salient and articulatorily clearer, while consonants, especially stop consonants, may be more difficult to do so. The current study focused on the word-final -er [-ɐ] and -t [-t] in German, the former of which was expected to have a higher degree of sonority than the latter of which. Both of them can be suffixal (e.g., Arbeit+er [aʁbaɪt+ɐ] "worker") and pseudo-suffixal (e.g., Vater [faːtɐ] "father") in German. The suffixal -er was found to be longer in duration and more clearly articulated than the pseudo-suffixal -er, while no such difference was found between the suffixal and pseudo-suffixal -t. These findings support the morphology-phonetics interaction and help to resolve the opposite findings regarding morphological effects on phonetic realizations.*

Keywords: Phonetics, morphology, sonority

1. Introduction

Morphological structures are not available to determine phonetic realizations (e.g., Levelt, Roelofs, and Meyer 1999). This assumption has been challenged by a number of studies that have reported effects of a morphological boundary, whether these morphological effects were found to be discrete or continuous/gradient in nature. For example, Walsh and Parker (1983) compared pairs of homophonous words, comparing presence and absence of a morphological boundary (e.g., *lapse* vs. *lap+s*, where "+" indicates a morphological boundary), and found a slightly longer duration for the morphemic /s/ (e.g., /s/ of *lap+s* compared to /s/ of *lapse*). Similar effects were later replicated by Seyfarth et al. (2017), finding longer stem and suffix durations for morphologically complex words (e.g., *free+s*) compared to morphologically simple words (e.g., *freeze*).

Hay (2007) extended the concept of the categorical morphological boundary effects (i.e, presence vs. absence of a morphological boundary) and found that the words with stronger morphological boundaries were associated with clearer phonetic realizations than those with weaker morphological boundaries, assuming that the words without a morphological boundary were located at the weakest end of this continuum of morphological boundary strength. Similar phonetic enhancement effects of a (gradient) morphological boundary were subsequently replicated by a number of studies such as Plag and Ben Hedia (2018), which found longer duration of the prefixes *un-* and *dis-* with higher segmentability (stronger morphological boundary) from their stems.

In contrast to these studies that reported phonetic enhancement effects of a morphological boundary, quite a few studies also reported the opposite effects of a morphological boundary, namely shorter duration associated with a stronger morphological boundary. Plag, Homann, and Kunter (2017) investigated the English word-final *-s* with and without a preceding morphological boundary and found longer duration for non-morphemic *-s*. Similar findings were also obtained by Zimmermann (2016) and Schmitz, Baer-Henney, and Plag (2021).

What causes these opposite findings regarding phonetic effects of a morphological boundary? One possible factor is the types of the items being investigated. Reduction effects of a morphological boundary have mainly been found for affixes consisting only of consonants such as English final /s/ and /z/ (Plag, Homann, and Kunter 2017; Zimmermann 2016; Schmitz, Baer-Henney, and Plag 2021). In contrast, enhancement effects have been found for the stem and the prefixes with a vowel in them (Hay 2007; Plag and Ben Hedia 2018; Seyfarth et al. 2017). In consistent with this tendency, vowels in prefixes have been found to be longer, while consonants in the same prefixes were shorter at the same time (Smith, Baker, and Hawkins 2012).

From the phonological perspective, vowels can be distinguished from consonants in terms of sonority. While sonority is a phonological concept, it has been suggested to have a certain phonetic basis (Clements 2009). Segments with higher sonority tend to have higher/better acoustic salience and also improved perceptibility. In other words, making clearer speech may have different degrees of effectiveness for different degrees of sonority. The speaker, implicitly knowing such differences, may choose to enhance those with higher sonority, rather than trying to make those with lower sonority more audible.

2. Methods

In order to test this possibility that the speaker selectively enhances those with high sonority because of their inherently better audibility, the current study focused on two German suffixes. One was *-er* [-ɐ], which could be a derivational suffix for a verb to indicate the agent of the action described by the verb (e.g., *Arbeit+er* [aʁbaɪt+ɐ] "worker") and could also be an inflectional comparative suffix for an adjective (e.g., *schön+er* [ʃøn+ɐ] "nicer/more beautiful"). The other suffix was *-t* [-t], which could be an inflectional suffix for third person singular or second person plural (e.g., *spiel+t* [ʃpiːl+t] "plays"). These two suffixes were chosen because they were both made of one segment and they were expected to be located at both ends of the sonority hierarchy (i.e., a vowel vs. a voiceless stop).

All the words with the word-final *-er* [-ɐ] and *-t* [-t] and their tongue position data were collected from the Karl Eberhards Corpus of spontaneously spoken southern German (Arnold and

Tomaschek 2016), regardless whether the word-final *-er* [-ɐ] and *-t* [-t] constituted genuine suffixes. Subsequently, each token was coded as to whether their word-final *-er* [-ɐ] or *-t* [-t] were genuine suffixes or pseudo-suffixes (i.e., a part of the stem). Those with the second person singular suffix *-st* [-st] were excluded to focus on the comparison of the suffixes of the length 1 (i.e., [-ɐ] vs. [-t]).

Suffix durations were calculated from time stamps available in the corpus, which marked the beginning and end of each segment. To capture phonological factors, each token was also coded as to whether they belonged to the words at the beginning or end of utterances (i.e., `UttInitial` and `UttFinal`). In addition, the number of syllables in each word, the number of syllables in each utterance, word duration, and utterance duration were calculated to represent speech rates (i.e., `NumSylWord`, `NumSylUtt`, `WordDur`, and `UttDur`). However, these last four phonological variables were correlated with each other. Therefore, they were combined by Principal Component Analysis, which showed the first principal component (i.e., `PC1`) alone already explained about 99% of the variance by the four phonological variables. Finally, to take predictability effects into account, word frequency was collected from the SdeWac corpus (Faaß and Eckart 2013) for each word. Word frequency as well as utterance-, word-, and suffix-durations showed skewed distributions and were log-transformed prior to the analysis.

For the duration data, log-transformed suffix duration (i.e., `SuffixDur`) was modeled as a function of log-transformed word frequency (i.e., `WordFreq`), `PC1`, `UttInitial`, `UttFinal`, `Speaker`, suffix identity (i.e., `Suffix`), and morphological status (i.e., `Morph`) in addition to the interaction between `Suffix` and `Morph`, using Generalized Additive Mixed-effects Models (GAMM; Wood 2017). `Suffix` contrasted *-er* and *-t* with *-er* as the reference level. `Morph` represented whether the word-final *-er* and *-t* constituted real suffixes. All the continuous variables (i.e., `WordFreq` and `PC1`) were modeled as smooth terms. `Speaker` was included as a random effect term.

Separately from the duration model, tongue tip positions during the target segments (i.e., [-ɐ] or [-t]) were modeled as a function of time (i.e., `Time`), with the covariates and factors mentioned above. `Time` was normalized between 0 and 1 for each token with the onset and offset of the segment/suffix in question being 0 and 1 respectively. In addition, `SuffixDur` was also included as an additional predictor, because longer duration allows speakers to make more dynamic movements of the tongue (Lindblom 1983; Tomaschek et al. 2018). Previous segments (i.e., `PrevSeg`), and next segments (i.e., `NextSeg`) were also taken into account as random effects. This was because tongue trajectories are influenced by adjacent segments especially at the beginning and the end of the target segment, namely coarticulation (Öhman 1966). `Morph` was allowed to interact with `Time` to estimate tongue trajectories for the pseudo-suffixal *-er*/*-t* and also to estimate differences in tongue trajectories between the pseudo-suffixal and suffixal *-er*/*-t*. The second term is called difference-curves (Baayen and Linke 2020). Because difference curves are represented by a single term, significance of the term would indicate significant differences between levels of the target factor variable (e.g., suffix vs. pseudo-suffix). With this model structure, two separate models were fitted for *-er* and *-t* (i.e., the `ER` model and the `T` model).

3. Results

For the duration data, word-final segments were estimated to be longer when they belonged to the words at the utterance-initial and final positions both, regardless of their morphological status ($\beta = 0.02, p < 0.01$ for `UttInitial` and $\beta = 0.39, p < 0.01$ for `UttFinal`). Effects of word frequency and speech rate were estimated somewhat U-shaped, though a majority of the data points ($\approx 99\%$) showed their greater values (i.e., higher frequency and faster speech rate) associated with shorter duration for both of the two variables. The suffix *-t* was estimated to be significantly shorter than *-er* ($\beta = -0.45, p < 0.01$). The suffixal *-er* was estimated to be significantly longer in duration than the pseudo-suffixal *-er* ($\beta = 0.06, p < 0.01$). The interaction between `Suffix` and `Morph` was significant ($\beta = -0.06, p < 0.01$) and indicated that the suffixal and pseudo-suffixal *-t* were not significantly different in duration from each other (Figure 1).

Table 1: Model summary for the duration data. Sfx=Suffix (*-er* vs. *-t*), Mor=Morph (pseudo-suffixes vs. genuine suffixes), T=TRUE.

	β	SE	t	p
Intercept	-2.32	0.01	-295.12	<0.01
Sfx=*-t*	-0.45	0.00	-126.09	<0.01
Mor=T	0.06	0.01	7.74	<0.01
UttInitial=T	0.02	0.00	4.26	<0.01
UttFinal=T	0.39	0.00	107.32	<0.01
Sfx=*-t*:Mor=T	-0.06	0.01	-7.36	<0.01
	edf	Ref.df	F	p
s(WordFreq)	1.95	2.00	177.71	<0.01
s(PC1)	1.98	2.00	51.22	<0.01
s(Speaker)	354.23	466.00	3.41	<0.01

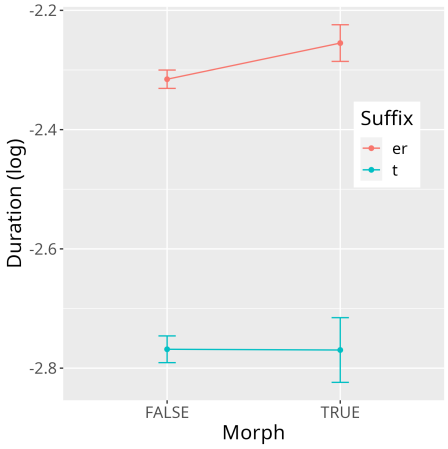


Figure 1: Predicted effects of `Morph` and `Suffix`.

For the tongue position data, the `ER` model predicted U-shaped tongue trajectories for *-er* with the suffixal *-er* significantly lower than the pseudo-suffixal *-er* ($\beta = -0.59, p < 0.01$). In addition, tongue trajectories were estimated significantly different between the suffixal and pseudo-suffixal *-er*, as shown in Figure 2. Figure 2 illustrates differences in tongue tip positions at different time points. The confidence intervals in

Figure 2 overlapping the horizontal line of $y = 0$ indicate no significant difference between the suffixal and pseudo-suffixal *-er* at that time point. According to Figure 2, significant differences are found in the middle of the segment *-er*, and the differences are negative, indicating that tongue trajectories of the suffixal *-er* are significantly lower than the pseudo-suffixal *-er* especially in the middle of the segment *-er*. None of the other control variables, namely `UttInitial`, `UttFinal`, `WordFreq`, and `PC1`, were significant.

Table 2: Model summary for the tongue position data of *-er*. Mor=Morph (pseudo-suffixes vs. genuine suffixes), T=TRUE.

	β	SE	t	p
Intercept	4.19	0.929	4.507	<0.01
Mor=T	-0.59	0.160	-3.702	<0.01
UttInitial=T	-0.02	0.115	-0.213	0.83
UttFinal=T	-0.89	0.957	-0.928	0.35
	edf	Ref.df	F	p
s(Time)	2.00	2.00	150.07	<0.01
s(Time):Mor=T	1.99	2.00	37.99	<0.01
s(WordFreq)	1.00	1.00	2.55	0.11
s(PC1)	1.68	1.90	1.03	0.36
s(PrevSeg)	20.14	23.00	714.94	0.57
s(NextSeg)	49.76	58.00	520.66	0.18
s(Speaker)	31.91	33.00	1620.28	0.04

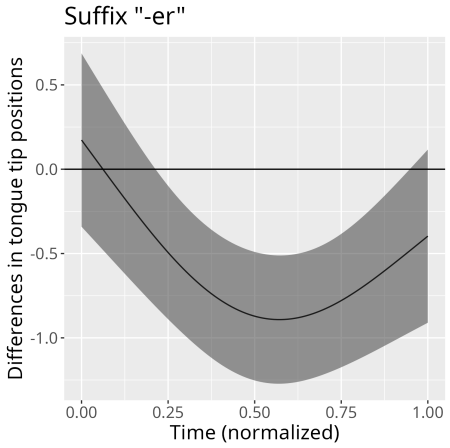


Figure 2: Predicted differences in tongue contours between the suffixal and pseudo-suffixal *-er*.

In contrast, there was no difference in tongue trajectories between the suffixal and pseudo-suffixal *-t*, as shown in Figure 3 ($\beta = 0.01, p \approx 0.91$ for the parametric term; $p \approx 0.45$ for the smooth term). Both of the suffixal and pseudo-suffixal *-t* showed upside-down U-shape tongue trajectories. `UttInitial` and `UttFinal` did not reach the significant level ($\beta = 0.09, p \approx 0.13$ for `UttInitial`; $\beta = -0.12, p \approx 0.85$ for `UttFinal`). Faster speech rate (i.e., `PC1`) was mainly associated with significantly higher tongue positions. Increase in word frequency was associated with higher tongue positions up to the middle frequency, after which increase in frequency showed lowering of tongue positions.

Table 3: Model summary for the tongue position data of *-t*. Mor=Morph (pseudo-suffixes vs. genuine suffixes), T=TRUE.

	β	SE	t	p
Intercept	8.27	0.89	9.29	<0.01
Mor=T	0.01	0.05	0.11	0.91
UttInitial=T	0.09	0.06	1.50	0.13
UttFinal=T	-0.12	0.67	-0.19	0.85
	edf	Ref.df	F	p
s(Time)	2.00	2.00	354.71	<0.01
s(Time):Mor=T	1.35	1.58	0.48	0.45
s(WordFreq)	1.99	2.00	35.91	<0.01
s(PC1)	1.98	2.00	21.89	<0.01
s(PrevSeg)	21.32	27.00	1197.37	<0.01
s(NextSeg)	81.96	102.00	135.11	<0.01
s(Speaker)	32.96	34.00	3428.02	<0.01

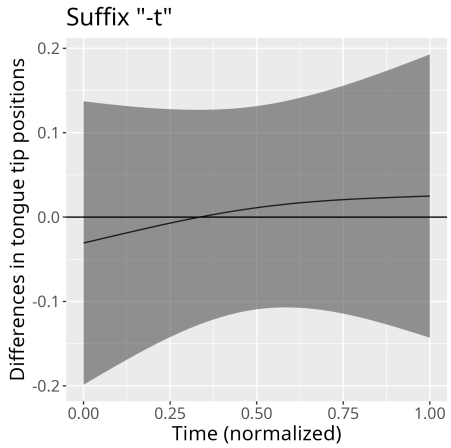


Figure 3: Predicted differences in tongue contours between suffixal and pseudo-suffixal *-t*.

4. Discussion

The current study looked into the possibility that phonetic enhancement/reduction effects of a morphological boundary were modulated by different degrees of sonority of the segment being investigated. To test this hypothesis, the current study investigated duration and tongue tip trajectories of *-er* and *-t*, both of which consisted of a single segment (i.e., [-ɐ] and [-t]) and can be suffixal or pseudo-suffixal.

The duration model predicted longer duration of the genuine suffix, compared to the pseudo-suffix. However, this effect of morphological status was found only for *-er*, and not for *-t*. The genuine and pseudo suffixes *-er* are realized as a vowel and therefore expected to be more sonorous than *-t*, namely a voiceless stop consonant. Therefore, these observations suggest that phonetic enhancement effects, at least in the context of morphological effects, are pronounced for segments of high sonority but diminished for those of low sonority.

The tongue position model for *-er* showed greater lowering of the tongue tip in the middle of *-er* when *-er* constituted the genuine suffix, compared to when it constituted the pseudo-suffix. Since *-er* is realized as a low open vowel, lower tongue positions indicate clearer articulation, namely phonetically enhanced realizations, which was found for the genuine suffix

compared to the pseudo-suffix. There was no such difference in articulation of the genuine and pseudo-suffixes of *-t*. These observations suggest that segments of high sonority (e.g., *-er*) are also articulatorily enhanced, while those of low sonority (e.g., *-t*) are not.

From the theoretical perspective, these results are not expected by a speech production model based on a modular-based feed-forward approach (e.g., Levelt, Roelofs, and Meyer 1999). In such a model, differences in morphological status are not relevant any more, once segments and metrical frames are retrieved for involved morphemes and combined into phonological words.

Why, then, do segments of high sonority get enhanced, while those of low sonority do not? Clements (2009) points out that the concept of sonority is associated with phonetic power relative to the weakest sound in the language being 1 (Fletcher 1929; Clements 2009). Those higher in the sonority scale tend to be more sonorant with greater phonetic power and retain a relatively low degree of acoustic loss (Clements 2009). In other words, those of high sonority are more persistent, while those of low sonority diminish faster. Because of greater degrees of acoustic loss, enhancing those with low sonority may not contribute to improve perceptibility so much as those with high sonority. The validity of this explanation remains open as an empirical question for future research.

From the practical perspective, the current findings help to understand the opposite findings regarding morphological effects on phonetic realizations. Some studies have found longer duration (e.g., Plag and Ben Hedia 2018), while others found shorter duration (e.g., Plag, Homann, and Kunter 2017), for affixes compared to pseudo-affixes. Most of the studies that found shorter duration for genuine suffixes compared to pseudo-suffixes are involved with consonants (e.g., *-s*). It is worth noting that some studies did found longer duration for the suffixal *-s* than the pseudo-suffix *-s* (Seyfarth et al. 2017; Walsh and Parker 1983). However, their findings are somewhat limited. Walsh and Parker (1983) found only 9 ms of differences between the genuine and pseudo *-s* without a statistical test performed on the difference. Seyfarth et al. (2017) found morphological effects on *-s* but not on *-d*. While the current findings showed that those of low sonority such as consonants are not easily enhanced, it remains as an empirical question whether they are ultimately enhanced, reduced, or completely insensitive to phonetic enhancement.

The current study indicated that phonetic realizations are influenced by degrees of sonority of segments in question, as well as their morphological status. Morphology, phonology, and phonetics are not separate processes but interacted with each other. Speech production models are called for that are able to accommodate such interactions.

5. Acknowledgements

The author would like to thank Dr. Jessie Nixon (University of Oldenburg) for helping him to improve the theoretical basis of this study.

6. 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 and Maja Linke (2020). "An introduction to the Generalized Additive Model". In: *A Practical Handbook of Corpus Linguistics*. Ed. by Magali Paquot and Stefan Th. Gries. Cham, Switzerland: Springer, pp. 563–591.

Clements, G. N. (2009). "Does Sonority Have a Phonetic Basis?" In: *Contemporary Views on Architecture and Representations in Phonology*. The MIT Press. DOI: 10.7551/mitpress/9780262182706.003.0007.

Faaß, Gertrud and Kerstin Eckart (2013). "SdeWaC: A corpus of parsable sentences from the web". In: *Language Processing and Knowledge in the Web*. Ed. by Iryna Gurevych, Chris Biemann, and Torsten Zesch. Darmstadt, Germany: Springer, pp. 61–68.

Fletcher, Harvey (1929). *Speech and Hearing*. New York, USA: D. Van Nostrand.

Hay, Jennifer (2007). "The phonetics of 'un'". In: *Lexical Creativity, Texts and Contexts*. Ed. by Judith Munat. Amsterdam/Philadelphia: John Benjamins, pp. 39–57.

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.

Öhman, S. E. G. (1966). "Coarticulation in VCV utterances: Spectrographic measurements". In: *The Journal of the Acoustical Society of America* 39, pp. 151–168. DOI: 10.1121/1.1909864.

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.

Schmitz, Dominic, Dinah Baer-Henney, and Ingo Plag (2021). "The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords". In: *Phonetica* 78.5-6, pp. 571–616. DOI: 10.1515/phon-2021-2013.

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.

Tomaschek, Fabian, Denis Arnold, Franziska Bröker, and R. Harald Baayen (2018). "Lexical frequency co-determines the speed-curvature relation in articulation". In: *Journal of Phonetics* 68, pp. 103–116.

Walsh, Thomas and Frank Parker (1983). "The duration of morphemic and non-morphemic /s/ in English". In: *Journal of Phonetics* 11.2, pp. 201–206.

Wood, Simon N. (2017). *Generalized additive models: An introduction with R*. 2nd. Boca Raton, Florida, USA: CRC Press.

Zimmermann, Julia (2016). "Morphological status and acoustic realization: Findings from New Zealand English". In: *Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology (SST-2016)*. December. Canberra: Australasian Speech Science and Technology Association (ASSTA), pp. 201–204.

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, Universidade Federal de Minas Gerais, Brazil*
[3]*Department of Electronic Engineering, Universidade Federal de Minas Gerais, Brazil*
[4]*MARCS Institute for Brain, Behavior & Development, Western Sydney University, Australia*
joao_vitor.possamai_de_menezes@tu-dresden.de

Abstract

Speech is a multimodal phenomenom at the perception and production ends, and that includes the suprasegmental level of speech. This paper focuses on the auditory-visual nature of lexical tones, a suprasegmental unit of speech that characterises tone languages. A multimodal corpus consisting of audio and Optotrak recordings of 33 markers in the face and head was recorded with 3 native speakers of Cantonese. The recorded trajectories of the Optotrak markers were parameterized as polynomial coefficients and used as input to Linear Discriminant Analysis models for classification between the 6 Cantonese lexical tones. Face and head motion were able to classify between lexical tones with above-chance accuracy for each speaker individually and for all speakers combined. Other analyses were carried out to determine which face regions and types of head motion had a stronger influence of the lexical tone classification accuracy, and the movement of the eyebrows and of the larynx stood out.

Keywords: lexical tones, Cantonese, auditory-visual speech, multimodal speech, Optotrak

1. 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 level (McGurk and MacDonald 1976). Later, the suprasegmental level of speech also became subject of studies on multimodality, and a series of visual correlates of speech prosody were found, such as eyebrow movement (Cave et al. 1996) and rigid body motion of the head (Yehia, Kuratate, and Vatikiotis-Bateson 2002). In this context, the multimodality of lexical tones, an element of speech prosody, 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. The first studies on the auditory-visual nature of lexical tone perception showed that, under certain circumstances, native speakers of Cantonese were able to differentiate between lexical tones with an above-chance accuracy based solely on visual stimuli (Burnham, Ciocca, and Stokes 2001). Additionally, the visual information of the speaker's lips during lexical tone production was found to aid speech perception under noisy conditions in Mandarin (Mixdorff, Hu, and Burnham 2005).

This first batch of studies relied on subjective perception experiments and were not designed to quantify the relevance of specific visual gestures, leaving as a gap the lack of quantitative evaluation of articulatory gestures. However, quantitative methods such as linear mixed models and computer vision gained relevance on the field in the following years. Burnham, Li, et al. (2019) used linear mixed models to analyse visual speech data from a Thai native speaker. Their results highlighted the relevance of larynx and head visual gestures for differentiating lexical tones. Garg et al. (2019) applied computer vision techniques on visual speech data of several Mandarin native speakers, and were able to suggest that specific Mandarin lexical tones are related with eyebrow and lip movements.

Our group has also performed qualitative analysis of acoustic and visual speech data, but with an approach based on statistical classification. The motivation for this approach is that the task of the classification model is similar to that of a speaker during a conversation: based on available information, decide what is being said. One advantage of this approach is the interpretability of some classification models, which allow the quantification of the relevance of specific visual gestures.

The present study is motivated by previous results from our group (Menezes et al. 2020; Burnham, Vatikiotis-Bateson, 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, Vatikiotis-Bateson, 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.

2. Methods

The speech production experiments for data acquisition were conducted at the MARCS Institute for Brain Behavior and De-

Table 1: *The units which compose the recorded corpus. The 12 word-units were each combined with the 6 cantonese lexical tones, resulting in* $12 * 6 = 72$ *units. The 8 syllable-units were each combined with the vowels /a, i, u/ and with the 6 cantonese lexical tones, resulting in* $8 * 3 * 6 = 144$ *units. The corpus was composed, therefore, by* $72 + 144 = 216$ *units.*

Word-units			Syllable-units	
ji	fɐn	jau	/p^h/	/k^h/
fu (x2)	jen	soei	/p/	/k/
si	hau	wai	/t^h/	/m/
sɛ	haːu		/t/	/n/

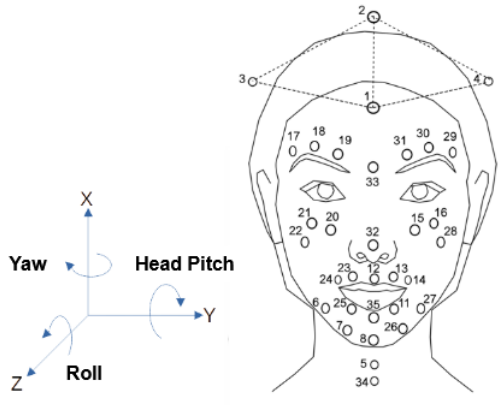


Figure 1: *Description of the Axes used throughout the experiments and the positions of the Optotrak marker's in the participant's face.*

velopment (Sydney, Australia) with 3 native speakers of Cantonese. The data set was composed by 216 isolated words in Cantonese, which were combinations of 36 phonetic strings with the 6 lexical tones, and its described in **Table 1**. The data set was recorded 4 times for each speaker, with the acoustic and visual speech data being recorded synchronously while the speakers produced the units in the data set. Each unit of the data set was manually segmented after recording, and processed individually.

The recorded visual speech data was measured with an NDI Optotrak Certus device (Northern Digital 2023), used to track the 3D (x, y, z) position of 33 active (LED emitting) 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. The acoustic speech data was captured by a high-quality microphone and then digitised and recorded by an NDI ODAU device at 44 100 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). The effect of this procedure is illustrated in **Figure 2**.

F0 contours were estimated based on the recorded acoustic speech data by the autocorrelation method implemented by Praat (Boersma and Weenink 2024). Each F0 contour was visually checked for errors, e.g., discontinuities or if any value was an octave above or below the speaker's regular F0. In such

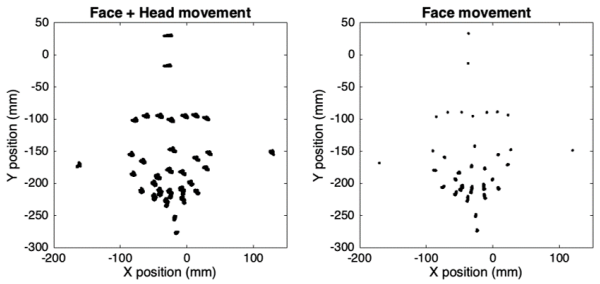


Figure 2: *Motion of the markers over 2000 samples (*33.33 s*). Left: original motion as captured by Optotrak consisting of both face and head components. Right: Face motion component after the head motion compensation procedure. Notice how the head motion is absent on the right pane.*

cases, the values of voicing threshold, silence threshold and octave cost of the Praat F0 pitch analysis were modified in order to obtain appropriate contours. In this work, the following lexical tone numbering convention is used, according to Chao (1930): tone 1 is high-level (55), tone 2 is rising (25), tone 3 is mid-level (33), tone 4 in low-falling (21), tone 5 is low-rising (23) and tone 6 is low-level (22).

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.

3. Results

The obtained results are presented in two subsections: the first on the overall lexical tone classification performance of each input domain, and the second on how different types of movements collaborated to the classification.

3.1. Classification accuracy

For each input domain (F0, head motion, face motion) and a concatenation of all input domains, classification performance was calculated as the average accuracy over 60 repetitions of 5-fold cross validated LDA models. **Table 2** presents the classification performances for each speaker individually and for all 3 speakers considered together.

Confusion matrices of the classification results for all participants are shown in **Figure 3** and provide a more detailed overview on how well each tone was classified.

3.2. Analysis of different types of movement

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. These analyses were performed with the data from all participants together, for higher generalization.

The rotation matrices of 2 LDA models (one using face mo-

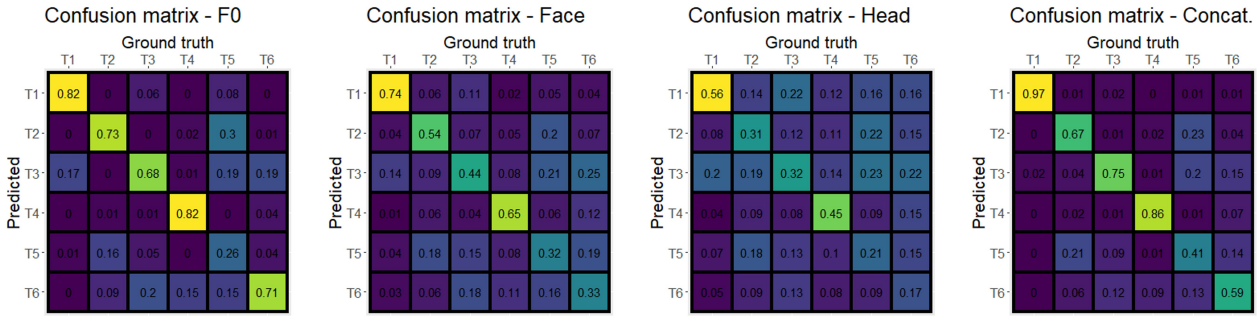


Figure 3: *Confusion matrices for the LDA models trained with input domains F0, Face, Head and their concatenation for all participants. The values within each cell are bounded between 0 and 1 and represent the percentage of predictions for each tone, with one tone as ground truth. The sum of each column of each matrix equals 1 (when the sum is less than 1 it is due to the rounding of each percentage to two decimal places).*

Table 2: *Mean and standard deviation of the classification performance for each input domain (and their concatenation) and each participant. Each participant is identified with a combination of letters to preserve its identity.*

Participant	F0	Face	Head	Concat.
CL	70.73% ±2.36%	57.82% ±3.21%	43.50% ±3.12%	66.64% ±3.14%
WMW	69.73% ±2.67%	61.14% ±3.16%	36.92% ±3.06%	69.52% ±3.14%
YL	83.94% ±2.36%	50.65% ±3.47%	36.69% ±3.39%	76.49% ±2.88%
All	66.94% ±1.67%	50.55% ±2.07%	33.85% ±1.91%	70.85% ±1.78%

tion 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 4** 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. **Table 3** shows how the classification accuracy varied with the complete removal of different face regions and head motion types in comparison to the results of **Table 2**. 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 most relevant signals in 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.

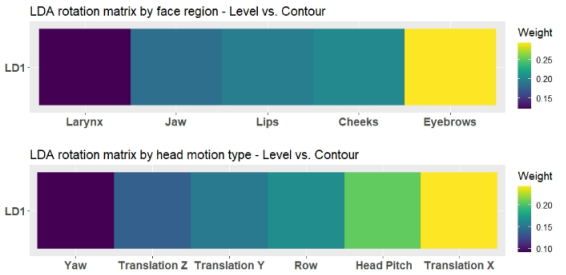


Figure 4: *Normalized heatmaps of face and head motion components and their weights in LDA rotation matrices differentiating Level vs. Contour tones.*

Table 3: *The absolute variations in mean accuracy of lexical tone classification when removing individual face regions and head motion types from the input data.*

Face region	Δ Acc.	Head motion type	Δ Acc.
Larynx	−7.83%	Yaw	+0.05%
Jaw	−0.66%	Head pitch	−1.28%
Lips	−1.48%	Row	−2.32%
Cheeks	−1.81%	Translation X	−1.94%
Eyebrows	−5.41%	Translation Y	−1.12%
		Translation Z	−4.72%

4. Discussion and conclusion

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. The multimodal input signal consisting of the concatenation of F0, Face and Head motion behaved differently for individual speaker and for all speaker together. In both scenarios, it achieved higher accuracies than the individual motion signals, but for individual speaker its accuracy was lower than that of the individual F0 signal, whereas for all speakers together it achieved the highest overall accuracy.

On one hand, the LDA algorithm is expected to lower its performance as the dimensionality of the input data grows, which explains the counter-intuitive lower accuracy achieved by the concatenation of all signals in comparison to F0. On the other hand, when the data from all speaker was considered together, the signal with most dimensions achieved the highest accuracy. This might have been enabled by higher generalisation in the training of the LDA model, with information of multiple speakers and multiple input signals.

The confusion matrices showed how well each tone was classified. For all input signals, tones 1 and 4 were classified with higher accuracy than other tones, whereas tone 5 was classified with the lowest accuracy. Comparing the results obtained with input domain F0 individually and the concatenated multimodal input, an increase in the classification accuracy of tones 1, 3, 4 and 5 and a decrease of it for tones 2 and 6 is observed. This suggests acoustic and motion signals convey information of varying relevance across tones, which might either interfere or collaborate with each other.

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. In Burnham, Vatikiotis-Bateson, et al. (2022), higher accuracy was achieved with head than with 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) and the present work when eyebrows were included corroborate this.

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, Vatikiotis-Bateson, 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.

5. References

Boersma, P. and D. Weenink (2024). *Praat: doing phonetics by computer [Computer program], Version 6.4.07, retrieved 17 March 2024 from http://www.praat.org/*.

Burnham, Denis, Valter Ciocca, and Stephanie Stokes (2001). "Auditory-visual perception of lexical tone". In: *Proceeding of the Eurospeech 2001*, pp. 395–398. DOI: 10.21437/Eurospeech.2001-63.

Burnham, Denis, Weicong Li, Chris Carignan, Virginie Attina, Benjawan Kasisopa, and Eric Vatikiotis-Bateson (2019). "Visual Correlates of Thai Lexical Tone Production: Motion of the Head, Eyebrows and Larynx?" In: *Proceedings of 15th International Conference on Auditory-Visual Speech Processing 2019*, pp. 69–72. DOI: 10.21437/AVSP.2019-14.

Burnham, Denis, Eric Vatikiotis-Bateson, Adriano Vilela Barbosa, João Vítor Menezes, Hani Camille 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: 10.1016/j.specom.2022.03.011.

Cave, C., I. Guaitella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser (1996). "About the relationship between eyebrow movements and Fo variations". In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*. Vol. 4, 2175–2178 vol.4. DOI: 10.1109/ICSLP.1996.607235.

Chao, Yuen Ren (1930). "A System of Tone-Letters". In: *Le Maître Phonétique* 45, pp. 24–27.

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: 10.1016/j.specom.2019.08.003.

McGurk, Harry and John MacDonald (1976). "Hearing lips and seeing voices". In: *Nature* 264, pp. 746–748. DOI: 10.1038/264746a0.

Menezes, João Vítor Possamai de, Maria Mendes Cantoni, Denis Burnham, and Adriano Vilela Barbosa (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.

Mixdorff, Hansjorg, Yu Hu, and Denis Burnham (2005). "Visual cues in Mandarin tone perception". In: *Proceeding of the Interspeech 2005*, pp. 405–408. DOI: 10.21437/Interspeech.2005-273.

Northern Digital, Inc. (2023). *Legacy products: NDI's 40-year history and transition*. URL: https://www.ndigital.com/products/legacy-products/.

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.

Yehia, Hani Camille, Takaaki Kuratate, and Eric Vatikiotis-Bateson (2002). "Linking facial animation, head motion and speech acoustics". In: *Journal of Phonetics* 30, pp. 555–568. DOI: 10.1006/jpho.2002.0165.

Yip, Moira (2002). *Tone*. Cambridge University Press. DOI: 10.1017/CBO9781139164559.

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

Abstract

This paper presents and discusses implications for the phonology-phonetics interface that can be drawn from three of our studies on coarticulatory development across childhood. Vocalic coarticulation towards the left (anticipatory) as well as towards the right side (carryover) was investigated in children between 3 and 9 years of age as well as in adults. To monitor horizontal tongue movements during pseudoword production, we used ultrasound tongue imaging. Our data provides evidence for a developmental decrease of coarticulation degree both in anticipatory as well as in carryover coarticulation that cannot easily be explained by look-ahead models positing complex translation mechanisms between discrete phonological and dynamic articulatory units. Instead, we argue that the developmental decrease as well as a finding of discontinuous vocalic anticipation in children can best be modeled within the coproduction framework, and therefore Articulatory Phonology.

Keywords: Coarticulation, speech and language acquisition, ultrasound, phonology-phonetics interface

1. 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. One big class of models assumes a look-ahead mechanism to scan intended utterances from left to right and change underlying segments based on their context. Daniloff & Hammarberg (1973), for example, explain coarticulation via phonological rules that spread binary features from right to left. 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 feature specification or orosensory space, respectively). While the look-ahead scanning mechanism, that is the basis of all three models, accounts for anticipatory coarticulation, it does not explain carryover effects. These coarticulatory effects from left to right are ascribed to purely mechanic-inertial aspects of speech production instead.

In contrast to pure look-ahead models, Articulatory Phonology (Browman & Goldstein 1986) is not built on translation rules between abstract and produced segments, but assumes articulatory gestures to underly both phonology and phonetics. In this framework, coarticulatory effects are not interpreted as an adjustment of ideal canonical segments to their context but as overlap between 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. Assuming language continuity (e.g., Fikkert, 2007), a good model of adult speech production must also be able to explain the developing system in children and in turn, studying child speech can help us model human speech production in general. 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 coarticulation degree. 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 degree 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 and carryover coarticulation results from motoric constraints, their evolution may differ greatly across childhood, while a common underlying mechanism like coproduction implies parallel development of the two coarticulatory directions.

2. Methods

The implications presented here are based on three studies: One on anticipatory coarticulation, one on carryover coarticulation, and one comparing the extent of anticipatory coarticulation in repeated versus read aloud speech. The first two studies work on the same data set for which 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 and processing 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/. After phonetic transcription and labeling with Praat (Boersma & Weenink, 2016), we semi-automatically detected the horizontal position of the highest point of the tongue dorsum in the ultrasound video frames of interest using custom-made MATLAB (2016) scripts as part of the SOLLAR platform. Only those productions rated to be segmentally correct and fluent were processed further. For anticipation, the ultrasound frames corresponding to the temporal mid- and endpoint of the article's /ə/ (schwa1_50, schwa1_100), and the temporal mid- and endpoint of C1 (C1_50, C1_100) were used. For carryover coarticulation, we looked at the temporal endpoint of the vowel (V_100), the temporal mid- and endpoint of C2 (C2_50,

C2_100), and the temporal midpoint of the final schwa (schwa2_50).

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 preceding (Noiray, Wieling, Abakarova, Rubertus, & Tiede, 2019) and following (Rubertus & Noiray, 2020) the acoustically defined interval of the vowel. By adding binary smooths, we were able to directly compare coarticulatory degree between age cohorts.

For the third study of interest, 32 additional children between 7 and 9 years of age and 16 adults were recorded within the same setup. In addition to repeating acoustically presented stimuli, they were asked to read aloud the pseudowords. Here, we focused on coarticulatory extent analyzing horizontal positions of the tongue dorsum's highest point at 23 vowel-preceding time points (at 0, 20, 40, 60, and 80 % of temporal length of each segment) and comparing the resulting movement trajectory over time between stimuli with front vowel /i/ and those with back vowel /u/ again using GAMMs (Rubertus, Popescu, & Noiray, n.d.) The point at which the smooth for /i/-stimuli and the one for /u/-stimuli start to diverge, i.e. when the tongue starts to move front for /i/ and back for /u/, is interpreted as the temporal onset of vowel anticipation. Adding a binary smooth in the GAMM, allowed us to directly compare coarticulatory extent between the reading and the repetition condition for children as well as for adults.

3. Results

3.1. Anticipatory and carryover coarticulation degree

The overarching result of our project is that coarticulatory degree substantially decreases across childhood, both in the anticipatory and the carryover direction. The youngest child cohort exhibited strongest vocalic coarticulation while adults' coarticulation was weakest. The contour plots in Figure 1 use color shades from yellow (back) to blue (front) to display the current horizontal position of the highest point of the tongue dorsum for each given position at vowel midpoint (y-axis) over time (x-axis). Here, only 3- and 7-year-olds' as well as adults' results for the /b/-context are shown. The complete results for anticipation can be found in Noiray et al. (2019) and those of carryover coarticulation in Rubertus & Noiray (2020). Close to the vowel (i.e., C1_100 for anticipatory and V_100 for carryover coarticulation), there are all different color shades reflecting a broad range of horizontal tongue positions. The further you move away from the vowel (towards the left for anticipatory and towards the right for carryover coarticulation), the less vowel-like the position gets. Importantly, this fan-like pattern is compressed for older age cohorts, indicating later movements towards vowel-specific positions and earlier movements back to central positions after the vowel, which means less vocalic coarticulation in both directions with increasing age. The comparisons between the age cohorts indicate a significant developmental change for both coarticulatory directions ($p < .01$).

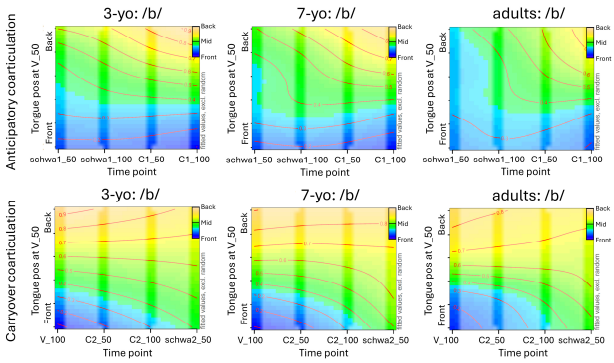


Figure 1. *Anticipatory (top row) and carryover coarticulation (bottom row) in /b/-context for 3- and 7-year-old children and adults. Color shades indicate the horizontal position of the tongue dorsum (yellow – back, blue – front) for a given position at vowel midpoint (y-axis) over time (x-axis).*

In addition to this main result, children exhibited discontinuous effects of carryover coarticulation in the /d/-context: During consonant production, i.e., at time points C2_50 and C2_100, children's tongue dorsum moves far forward for a broad range of preceding vowel positions as indicated by the high blue portion in 3-year-olds' plot in Figure 2. At schwa2_50, however, the tongue dorsum's position is more vowel-like again. In adults (right-hand side of Figure 2), the forward movement is not as pronounced as in children and there is less vowel-dependence during the final schwa.

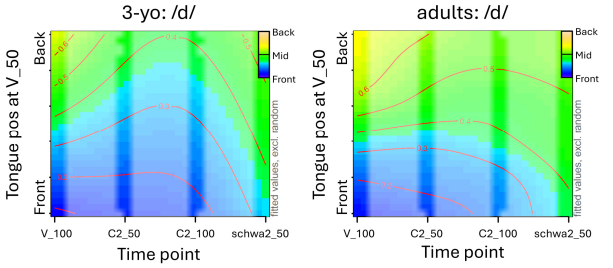


Figure 2. *Carryover coarticulation in /d/-context for 3-year-olds and adults. Color shades indicate the horizontal position of the tongue dorsum (yellow – back, blue – front) for a given position at vowel midpoint (y-axis) over time (x-axis).*

3.2. Coarticulatory extent in reading versus repetition

Results of the comparison between read aloud and repeated speech are displayed in Figure 3 that plots tongue positions over time separately per cohort (children – left, adults – right) and condition (reading – top, repetition – bottom). Front tongue dorsum positions are indicated by low y-values and back positions by high y-values for /i/- (in blue) and /u/-stimuli (in pink). At the beginning of the utterances, the tongue trajectories for /i/- and /u/-stimuli are very similar within each plot, the time point at which they start to diverge, however, differs: Beginning readers (children) exhibited limited coarticulatory extent when reading aloud compared to repeating stimuli, while proficient readers (adults) did not differ in coarticulation extent between these modalities. The time windows of significant difference between /i/ and /u/ are highlighted in red. While the difference in coarticulatory extent between the reading and the repetition condition is significant in children ($p = .016$), it is not in adults ($p = .588$).

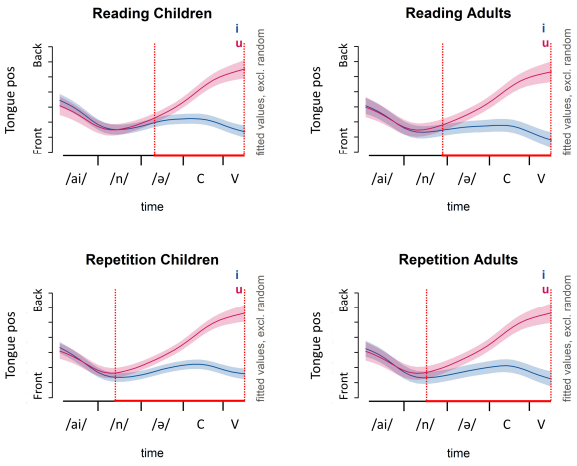


Figure 3. *Results of the comparison between read aloud and repeated speech. Non-linear smooths for /i/- (blue) and /u/-stimuli (pink) for the position of the highest point of the tongue dorsum (0 – front, 1 – back) over time. The corresponding acoustic segment boundaries are indicated on the x-axes. Significant time windows are indicated in red. Children – left, adults – right. Reading – top, Repetition – bottom.*

4. Discussion and conclusion

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. Within these frameworks, a developmental increase of coarticulatory degree would be expected (e.g., Redford, 2019). The coproduction framework (Fowler, 1980), instead, provides a plausible explanation for the developmental decrease: 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 supported by the parallel developments in anticipatory and carryover coarticulation highlighted in our studies. Here, children's stronger vocalic coarticulation is envisioned as broader overlap of vocalic with surrounding segments' gestures than in adults. The high prominence of stressed vowels in children's input is well-known and there is evidence that they serve as anchors both in perception as well as production (Cutler & Mehler, 1993; Fox & Dodd, 1999; Höhle, Bijeljac-Babic, Herold, Weissenborn, & Nazzi, 2009). Children's limited inhibition capacity may be one reason for their extraordinary strong activation of the stressed vowel (Bjorklund & Harnishfeger, 1990; Tilsen, 2013). The observed discontinuous vocalic effects provide further 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 gradual compression of vocalic activation curves progressively limiting vocalic overlap (cf. Nittrouer 1993; see Figure 4).

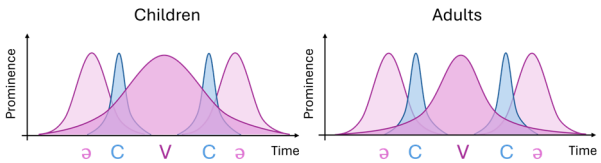


Figure 4. *Segments' hypothesized prominence over time in utterances of the form əCVCə, following Nittrouer (1993, p. 961).*

Literacy acquisition may be one among potentially many factors stimulating a decrease in the width of vocalic activation. Indeed, exposition to alphabetic orthographies raises awareness of the composite nature of speech and of representational units at the phonemic level (e.g., Goswami, 2000), which may sharpen the borders to surrounding speech segments both on a representational as well as on the articulatory level. In addition, orthographies like German that do not graphically emphasize stressed vowels, may contribute to reducing the relative prominence of stressed vowels by conveying the impression of equivalence. We intend to pursue this work further in future empirical investigations.

To conclude, our data provides evidence for a decrease of lingual vocalic coarticulation across childhood. Even in segmentally correct and fluent productions, children's speech still differs from adults' in fine phonetic details. The empirical data can best be modelled by a developmental compression of vocalic overlap within the coproduction model and therefore lends support to the framework of Articulatory Phonology highlighting the close connection between abstract phonology and speech production. As promoted by Vihman and Croft (2007), Redford (2019), and others, our work shows 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. Further results and in-depth discussion of the implications that our work on coarticulatory patterns in childhood bears for speech motor and phonological development can be found in Rubertus (2024).

5. Acknowledgements

This research was supported by two grants from the Deutsche Forschungsgemeinschaft (255676067 and 1098, recipient: Aude Noiray) and the Marie Skłodowska-Curie Innovative Training Network under Grant 641858. We thank the LOLA-lab team for their indefatigable support and all our participants for letting us watch their tongues and making this research possible.

6. References

Bjorklund, D. F., & Harnishfeger, K. K. (1990). The resources construct in cognitive development: Diverse sources of evidence and a theory of inefficient inhibition. *Developmental Review*, *10*(1), 48–71. doi: 10.1016/0273-2297(90)90004-N

Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by computer (Version 6.0.20) [Computer program]. Available from http://www.praat.org/.*

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*(1986), 219–252. doi: 10.1017/s0952675700000658

Cutler, A., & Mehler, J. (1993). The periodicity bias. *Journal of Phonetics*, *21*, 101–108.

Daniloff, R. G., & Hammarberg, R. E. (1973). On defining coarticulation. *Journal of Phonetics*, *1*(3), 239–248. doi:

10.1016/s0095-4470(19)31388-9
Fikkert, P. (2007). Acquiring phonology. *Handbook of Phonological Theory*, 537–554.
Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*(1), 113–133. doi: 10.1016/S0095-4470(19)31446-9
Fox, A. V, & Dodd, B. J. (1999). Der Erwerb des phonologischen Systems in der deutschen Sprache [The phonological acquisition of German]. *Sprache-Stimme-Gehör*, *23*(4), 183.
Goswami, U. (2000). Phonological representations, reading development and dyslexia: Towards a cross-linguistic theoretical framework. *Dyslexia*, *6*(2), 133–151.
Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*(3), 594.
Höhle, B., Bijeljac-Babic, R., Herold, B., Weissenborn, J., & Nazzi, T. (2009). Language specific prosodic preferences during the first half year of life: Evidence from German and French infants. *Infant Behavior and Development*, *32*(3), 262–274. doi: 10.1016/j.infbeh.2009.03.004
Keating, P. A. (1988). The window model of coarticulation: articulatory evidence. *UCLA Working Papers in Phonetics*, *69*, 3–29.
MATLAB. (2016). *MATLAB and Statistics Toolbox Release, The MathWorks, Inc., Nattick, Massachusetts, United States.*
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. doi: 10.1044/jshr.3605.959
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. doi: http://doi.org/10.5334/labphon.241
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. doi: 10.1044/2019_JSLHR-S-CSMC7-18-0208
Redford, M. A. (2019). Speech production from a developmental perspective. *Journal of Speech, Language, and Hearing Research*, *62*(8S), 2946–2962. doi: 10.1044/2019_JSLHR-S-CSMC7-18-0130
Rubertus, E. (2024). *Coarticulatory changes across childhood: implications for speech motor and phonological development* ((Doctoral Thesis). University of Potsdam, Germany). doi: https://doi.org/10.25932/publishup-63012
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. doi: http://doi.org/10.5334/labphon.228
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*.
Tilsen, S. (2013). A dynamical model of hierarchical selection and coordination in speech planning. *PloS One*, *8*(4), e62800.
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. doi: 10.1123/mcj.15.1.118

Laterals in simplex vs. complex syllable codas: a comparison of four languages

Anisia Popescu[1], Ioana Chitoran[2]

[1]*LISN, Université Paris Saclay, France*
[2]*CLILLAC-ARP, Université Paris Cité, France*
anisia.popescu@universite-paris-saclay.fr

Abstract

English dark /l/ in coda clusters has been shown to exhibit non-local coordination patterns (i.e., vowel shortening when going from singleton to complex codas - e.g., gull-gulp). The unpredicted organization has been attributed to the articulatory complexity and timing of the coda dark /l/. The present study further investigates this issue by analyzing coda coordination patterns as a function of /l/ darkness in Russian, English, Romanian and Georgian, four languages that differ in the gestural synergies of the coda lateral. The acoustic analysis shows that while predictions on coda coordination patterns based on the gestural composition of the /l/ are confirmed for Russian, English and Romanian, they are not confirmed in the case of Georgian.

Keywords: speech production, lateral allophony, coordination patterns

1. Introduction

Syllable level coordination patterns, and in particular the asymmetry between onsets and codas, first described within the framework of Articulatory Phonology by Browman and Goldstein (1988) have been the focus of many studies. Within this framework, while onsets are hypothesized to have global coordination patterns, codas are timed locally. A global coordination implies that consonant gestures are synchronously timed with the following vowel gesture, triggering articulatory rearrangements with increasing phonotactic complexity. Local coordination patterns do not trigger rearrangements, the vowel and consonant gestures are sequentially timed (i.e., one gesture starts after the previous one is deactivated). Since the seminal paper of Browman and Goldstein (1988), quite a few studies have looked at cross-linguistic differences in onset coordination patterns (Honorof and Browman 1995; Brunner, Geng, and Sotiropoulou 2014; Pouplier 2012; Hermes, Grice, et al. 2008; Hermes, Ridouane, et al. 2011; Shaw et al. 2009), identifying both language- and consonant-cluster-specific patterns. Far fewer studies have looked at between-language differences of coordination patterns in codas (Marin and Pouplier 2014). One particular study, Marin and Pouplier (2010), identified a non-local organization pattern in coda position. This articulatory-acoustic study of American English onset and coda clusters revealed that in lateral coda clusters the presence of the lateral triggers a shift of the /l/ gestures towards the preceding vowel, similar to patterns found in complex onsets. Acoustically, a shortening of the vowel is observed. The authors attribute this divergent pattern to perceptual consequences of the coda /l/'s overlap with the vowel. In an acoustic study, Katz (2012) replicated and extended their finding to coda rhotics, as well. In a subsequent study, Marin and Pouplier (2014) did not find the divergent pattern for German and Romanian coda /l/, but replicated it for Romanian trills. These results led the authors to amend their original hypothesis of attributing the non-local coda coordination to perceptual recoverability, suggesting that it is more likely due to the articulatory characteristics of the liquid. Coda liquids in American English (dark /l/ and rhotic) are produced with a double lingual gesture: a vocalic gesture - tongue dorsum (TD) retraction for the lateral and tongue root (TR) retraction for the rhotic - that precedes a consonantal tongue tip (TT) gesture. Similar to the English dark /l/, the rhotic trill in Romanian is also produced with a double gesture: a vocalic TR gesture that precedes and acts as an anchor for the consonantal TT trilling gesture. The English liquids and the Romanian /r/ (the cases where coda global organization was found) all involve the presence of a double lingual gesture and an earlier occurring vocalic gesture.

In this paper we therefore hypothesize that non-local organization in coda position occurs due to the presence of an earlier vocalic gesture that triggers gestural competition between the vowel nucleus and the vocalic gesture of the liquid, resulting in a reorganization of articulatory synergies in the rime.

To test this hypothesis, we focus on lateral consonants by comparing coda coordination patterns, acoustically, in four languages that differ in the gestural synergies of their coda lateral consonant: Russian (coda dark /l/ - Recasens (2012)), English (coda dark /l/ - Sproat and Fujumura (1993)), Georgian (clear - dark /l/ allophony based on vowel frontness – (Robins and Waterson 1952; Chigogidze 2011)) and Romanian (coda clear /l/ - Recasens (2012)). Robins and Waterson (1952) describe the Georgian lateral allophony as "decided “dark” velarized quality in all positions except before /e/ and /i/" (page 63), suggesting that Georgian /l/ is always dark /l/ in codas. Their study is, however, based on a single informant. The data in the present study suggests that both clear and dark /l/ varieties are present in coda position. Unlike dark /l/, clear /l/ lacks an earlier TD retraction gesture (Sproat and Fujumura 1993; Narayanan, Alwan, and Haker 1997) and is therefore not expected to trigger a non-local organization in coda position.

We test our hypothesis using acoustic data. The acoustic effect of global coordination patterns in coda position is a shortening of the vowel in cluster tokens compared to their respective singleton counterpart.

2. Methods

A total of 23 native speakers - Russian (5), American English (6), Georgian (6) and Romanian (6) - were recorded producing three repetitions of target singleton-cluster pairs (C)CVL-(C)CVLC with varying front/back vowel contexts, embedded in a carrier phrase in each respective language. Target words were hand segmented and labeled in Praat (Boersma and Weenink (2022)) based on the waveform and wide-band spectrograms.

The syllable rhyme was segmented for each target word. The onset of the vowel was marked at the onset of F1. The end of the visible formant structure on the spectrogram corresponded to the end of the postvocalic /l/. Given that the boundary between vowel and lateral (especially for dark /l/) was not always robustly identifiable, two different duration measures were considered for all the data: vowel + lateral (VL) sequences, and the interval between the midpoint of the vowel and the midpoint of the lateral (V50-L50) following Durvasula (2023). Raw duration measures were normalized dividing each duration measure by the articulation rate, calculated as the number of phones per second. To measure the acoustic shortening degree in cluster vs. singleton tokens, we defined a duration ratio as the ratio between each cluster and the corresponding singleton token (e.g. for the *sill [sɪɬ]- silk [sɪɬk]* pair the duration ratio is $\frac{\text{ɪl- duration}_{\text{sɪlk}}}{\text{ɪl- duration}_{\text{sɪl}}}$). 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 (i) the mean standard deviation of the duration ratios found in our data, (ii) the lowest standard deviation found in our data (i.e., Russian) and (iii) the highest standard deviation found in our data (i.e., Georgian). Testing H with three different standard deviations ensures the choice of standard deviation does not bias the results.

Formant values (F1, F2, F3) were extracted at the lateral midpoint (when the vowel-lateral boundary was identifiable) or at the steady state of the lateral (when the vowel-lateral boundary wasn't robustly identifiable). The darkness degree was calculated as the difference between the F2 and F1 values (/l/$_{\text{darkness}}$= F2 - F1).

2.1. Statistical analysis

The degree of /l/ darkness and the duration ratios between singleton and cluster tokens were used as dependent variables in linear mixed effects models (*lme4* - (Bates et al. 2015)). For all three models (/l/ darkness and two duration measures) predictors included *Language* (H (reference level), Russian, English, Georgian, Romanian), *Vowel Quality* (front, back), and *Sex* (F, M) as fixed factors and *Participant* and *Repetition* as random effects with random intercepts. An interaction term between *Language* and *Vowel Quality* was included.

2.2. Predictions

We expect shortening of VL and V50-L50 duration between clusters and singletons in Russian, English, and Georgian rimes with back vowels (dark /l/ in coda). No shortening is expected in Romanian and front vowel Georgian rimes (clear /l/ in coda).

3. Results

Results will be presented in two stages. First we compare lateral darkness across the four languages, followed by the results on acoustic shortening in the rime.

3.1. Coda /l/ in four languages

We first present acoustic differences of coda lateral allophones in the four languages considered: Russian, English, Georgian and Romanian. The measure of /l/ darkness is the difference between the second and first formants (F2-F1). Clear /l/ is characterized by higher, and dark /l/ by lower values of F2-F1.

The descriptive analysis shows that the degree of lateral darkness is a gradient feature across languages. Russian has the darkest lateral of the four languages (overall mean$_{\text{F2-F1}} = 403$). English has the second darkest lateral (overall mean$_{\text{F2-F1}} = 538$). The third darkest lateral in our data (independent of vowel context) is the Georgian one (overall mean$_{\text{F2-F1}} = 950$). Finally, as expected, Romanian has the clearest coda lateral ((overall mean$_{\text{F2-F1}} = 1248$).

Table 1: *Mean F2-F1 values of /l/ in front and back vowel context per language and participant. The absolute value of the difference between front and back vowel contexts is displayed in the rightmost column.*

Language	Speaker	Front V	Back V	|Δ|
Russian	RU01	249	378	129
	RU02	482	247	235
	RU03	524	400	124
	RU04	440	448	8
	RU05	476	505	29
English	EN01	370	531	165
	EN02	696	531	165
	EN03	756	601	155
	EN04	670	500	170
	EN05	684	487	197
	EN06	341	331	10
Georgian	GE01	1489	808	681
	GE02	881	735	146
	GE03	649	640	9
	GE04	1275	802	473
	GE05	1203	706	497
	GE06	1528	666	862
Romanian	RO01	1018	1068	50
	RO02	1466	1459	7
	RO03	1590	1366	224
	RO04	1315	1299	16
	RO05	1107	1131	24
	RO06	1200	1073	126

A certain degree of inter-speaker and vowel context variability is observed. Table 1 shows the mean values of /l/ darkness for all participants as a function of front and back vowels. The absolute value of the difference between front and back vowel /l/ darkness measures is displayed in the rightmost column (Δ = |F2-F1$_{\text{front V}}$ - F2 - F1$_{\text{back V}}$|). Differences in mean values of coda /l/ darkness between front and back vowel contexts are non-zero for all the languages considered, suggesting that some degree of coarticulation is present. Figure 1 illustrates the distribution of /l/ darkness measures for all languages as a function of vowel context.

The linear mixed model shows that Russian and Romanian do not exhibit a vowel context allophony. No significant difference in F2-F1 is found between front and back vowel contexts (Russian: Est. $\sim$ 31.54, t-value $\sim$ 0.63, p-value $\sim$ 0.52; Romanian: Est. $\sim$ 54.2, t-value $\sim$ 1.4, p-value $\sim$ 0.14). Both English and Georgian show significant differences in F2-F1 values depending on vowel context, with coda /l/ being clearer after a front vowel than after a back vowel. The effect is larger for Georgian than for English (Georgian: Est. $\sim$ 442, t-value $\sim$ 16.01, p-value < 0.001; English: Est. $\sim$ 119, t-value $\sim$ 3.37, p-value < 0.001).

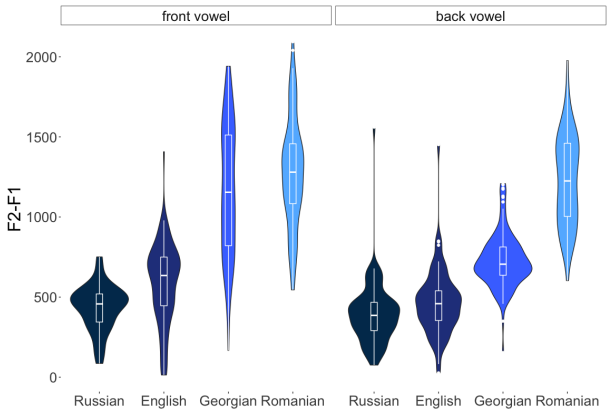


Figure 1: *F2-F1 values as a function of language (Russian, English, Georgian, Romanian) and vowel quality (front vs. back).*

3.2. Acoustic shortening in singleton-cluster pairs

In this section we present acoustic differences of vowel-lateral shortening between cluster and singleton in the four languages (Russian, English, Georgian, Romanian) compared to a hypothetical language H that has no shortening. Figure 2 illustrates the normalized VL duration ratios as a function of language and vowel quality.

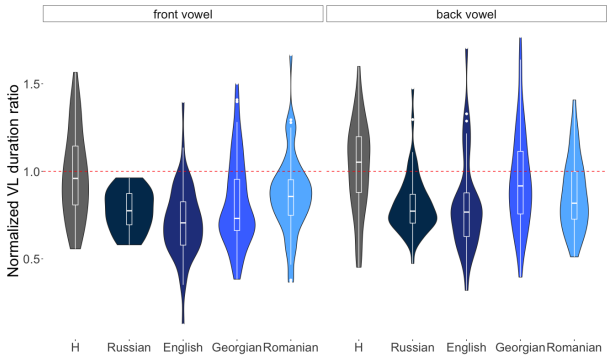


Figure 2: *Normalized ratios of vowel-lateral (VL) duration as a function of language (H, Russian, English, Georgian, Romanian) and vowel quality (front vs. back). The red horizontal line indicates a ratio of 1 (i.e., VL duration in singleton = VL duration in cluster token)*

Results of the linear mixed models only partially confirm our prediction, and cross-measure differences (VL vs. V50-L50) are found. The VL duration ratio results show that, as expected, Russian and English both have significantly higher degrees of shortening, while Romanian shows no differences in shortening compared to language H in either front or 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). Results for the VL duration ratio are summarized in Table 2.

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 language H only in front vowel contexts). The V50-L50 measure is, however, the less reliable one in our case because of the difficulty of identifying a precise acoustic boundary between the vowel and the lateral coda, especially in back vowel - dark /l/ and front vowel – clear /l/ sequences.

Table 2: *Model results of VL duration ratios as a function of Language and Vowel Context. The reference level for language is the hypothetical language H (normal distribution of mean=1 and standard deviation equal to the mean standard deviation found in our data).*

	Front V			Back V		
Lang.	Est.	t	p.	Est.	t	p
RU	-0.21	-2.5	< 0.01	-0.23	-3.27	< 0.01
EN	-0.26	-3.9	< 0.001	-0.20	-2.91	< 0.01
GE	-0.17	-2.65	< 0.05	-0.09	-1.4	0.16
RO	-0.10	-1.45	0.15	-0.15	-1.9	0.05

4. Discussion and conclusion

The present study set out to test 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, which shows the reverse pattern from the one predicted. One possible explanation for the unexpected pattern is that the degree of darkness could play a role. In our data, Georgian dark /l/ is significantly less dark than Russian and English dark /l/.

An effect of vowel context was found for both English and Georgian. We expected it for Georgian, which has a reported vowel context allophony in onsets. Previous literature on Georgian suggested dark /l/ in codas, but our data reveals much more variability. Data from additional speakers, as well as an analysis of individual speaker data, may reveal whether this difference can be attributed to coarticulatory effects, or to allophony.

Depending on duration measure (VL vs. V50-L50), significance levels change. Given the difficulty of robustly identifying the boundary between vowel and lateral in some cases, we used a proxy measure consisting of the vowel-lateral (VL) sequences. We acknowledge that other changes may occur within this VL interval, which require further careful acoustic and articulatory investigation. At the same time, V50-L50 is not a robust measure. While it works well for obstruents, it is less efficient for sonorants, since vowels are less easily separable from sonorants acoustically.

The present study relies only on acoustic data to confirm predictions derived from theoretical articulatory characteristics of laterals. In order to better understand the relationship between /l/ darkness and coda coordination patterns, articulatory data will be collected to precisely compare the timing of the articulatory gestures in the lateral coda rime.

5. Acknowledgements

The authors thank two anonymous reviewers for their useful comments, and express their gratitude to all the native speakers who participated in the study. The study received funding from the Labex EFL (ANR-10-LABX-0083-LabEx EFL) to Université Paris Cité.

6. References

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). "Fitting linear mixed-effects models using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48.

Boersma, P. and D. Weenink (2022). "Praat: doing phonetic by computer". In: *From http://www.praat.org/*.

Browman, C. and L. Goldstein (1988). "Some notes on syllable structure in articulatory phonology". In: *Phonetica* 45.2-4, pp. 140–155.

Brunner, J., C. Geng, and A. Sotiropoulou S.and Gafos (2014). "Timing of German onset and word boundary clusters". In: *Journal of Laboratory Phonology* 5.4, pp. 403–454.

Chigogidze, A. (2011). *On the Interaction of Syntax and Phonology in Georgian*. Master's thesis.

Durvasula, K. (2023). "A simple acoustic measure of onset complexity". In: *Proceedings of ICPhS 2023*, pp. 2010–2014.

Hermes, A., M. Grice, D. Mücke, and H. Niemann (2008). "Articulatory indicators of syllable affiliation in word initial consonant clusters in Italian". In: *In Proceedings of the 8th ISSP*, pp. 433–436.

Hermes, A., R. Ridouane, D. Mücke, and M. Grice (2011). "Kinematics of syllable structure in Tashlhiyt Berber : the case of vocalic and consonantal nuclei". In: *In Proceedings of the 9th ISSP*, pp. 401–408.

Honorof, D. and C. Browman (1995). "The center or edge: How are consonant clusters organized with respect to the vowe?" In: *In Proceedings of the 13th ICPhS*, pp. 552–555.

Katz, J. (2012). "Compression effects in English". In: *Journal of Phonetics* 40.3, pp. 390–402.

Marin, S. and M. Pouplier (2010). "Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model". In: *Motor Control* 14.3, pp. 380–407.

— (2014). "Articulatory synergies in the temporal organization of liquid clusters in Romanian". In: *Journal of Phonetics* 42, pp. 24–26.

Narayanan, S. S., A. A. Alwan, and K. Haker (1997). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I: The laterals". In: *Journal of the Acoustical Society of America* 101.2, pp. 1064–1077.

Pouplier, M. (2012). "The gestural approach to syllable structure: universal, language- and cluster-specific aspects". In: *Speech planning and dynamics*. Berlin: Peter Lang.

Recasens, D. (2012). "A cross-language acoustic study of initial and final allophones of /l/". In: *Speech Communication* 54.3, pp. 368–283.

Robins, R.H. and N. Waterson (1952). "Notes on the phonetics of the Georgian word". In: *Bulletin of the School of Oriental and African Studies* 141, pp. 55–72.

Shaw, J., A. Gafos, P. Hoole, and C. Zeroual (2009). "Temporal evidence for syllabic structure in Moroccan Arabic: data and model". In: *Phonology* 26, pp. 187–215.

Sproat, R. and O. Fujumura (1993). "Allophonic variation in Enlish /l/ and its implications for phonetic implementation". In: *Journal of Phonetics* 21.2, pp. 291–311.

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

Abstract

In this study involving ten participants, we simultaneously recorded acoustic and articulatory data using electromagnetic articulography (EMA) to investigate lexical stress production in L2 English among Taiwanese Mandarin (TWM) speakers. Analysis of dynamic time series data uncovered hyper-articulation of lingual articulators and the jaw in stressed syllables, alongside distinct tonal contours related to stress and syllable position. Furthermore, a gestural analysis revealed longer consonant gesture plateaus and longer CV lag in stressed syllables but no difference in gesture velocity.

Keywords: L2 English, Lexical stress, Taiwanese Mandarin, EMA

1. Introduction

Taiwanese Mandarin (TWM) is an East Asian tone language that lacks discernible word-level prominence. This study investigates the articulatory and acoustic correlates of stressed and unstressed syllables in L2 English as produced by native TWM speakers. While there have been similar studies, such as Kim's (2021) study involving speakers of Beijing Mandarin (BJM) and Shanghai Mandarin (SHM), our research departs substantially in its methodology and focus. We focus on TWM, a dialect which exhibits very limited use of the neutral tone—often cited as evidence of a trochaic foot in BJM. Crucially, our methodology differs by analysing dynamic trajectories of articulators over time through generalised additive mixed modelling (GAMM), moving beyond the static midpoint measurements used in earlier studies. In addition, we present an analysis of gestural duration and CV lag.

In L1 English, stressed syllables are described as 'hyper-articulated' (de Jong, 1995), while in Greek, they are said to involve 'longer, larger, and faster gestures than their unstressed counterparts' (Katsika and Tsai 2021). Acoustically, the key components used to distinguish stress in L1 English include intensity and duration (Fry 1955), vowel reduction (Delattre 1969), and fundamental frequency (F0) (Lieberman 1960) (although Pierrehumbert 1980 argues that F0 is related to pitch accent rather than being a direct correlate of stress). This exploratory study aims to determine which of the above articulatory and acoustic correlates, implicated in native stress production, are used by L1 TWM speakers in their L2 English production.

2. Methods

2.1. Participants, stimuli and recording procedures

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 focused 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 completed ten repetitions of each word, whereas the remaining two could only complete seven repetitions of each due to time constraints. 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), LA (lip aperture - the euclidean distance between the upper and lower lip senors) and JAW (lower incisor). Acoustic data were recorded simultaneously at 24 kHz.

2.2. Articulatory measurements

Articulatory measurements were made in Matlab using Mview (Tiede 2005). For the vowel analysis, vocalic portions of the articulatory trajectories were identified using the acoustic data as a guide. For the gestural duration and CV lag analyses, articulatory gestures were identified using the *findgest* algorithm in Mview, which identifies gestural landmarks based on a peak velocity threshold of 20%. The following sensors were used to measure the syllable-initial consonant gestures: TDz (where 'z' indicates vertical movement) for [k]on, TBz for [dʒ]est and [dʒ]ect, TTz for [d]i and LA for [f]lict and [p]ro. For the vowel in the CV lag analysis, TDz was used for d[aɪ]. The gestural plateau, specifically the hold phase, was defined as the NOFFS (nucleus offset) timestamp minus the NONS (nucleus onset) timestamp, as shown in Figure 1. CV lag was defined as the interval between the NONS of the consonant and the NONS of the vowel. To control for speech rate, gesture durations were time normalised using an anchor point in the following word. In addition to the duration of gestural plateaus, peak velocity (towards the closure, i.e. sensor velocity at the PVEL (peak velocity) timestamp in Figure 1) and amplitude normalised peak velocity (stiffness; see Roon et al. 2021) were also measured.

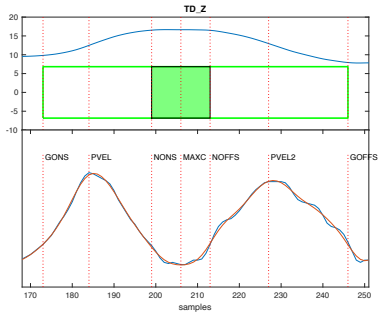


Figure 1: *Example of a [k] gesture—Upper panel: Vertical movement of tongue dorsum sensor; Lower panel: Velocity of sensor*

2.3. Acoustic measurements

The acoustic study investigated the realisation of stressed versus unstressed syllables in terms of three suprasegmental dimensions: intensity (dB), duration (ms), and F0 (Hz) and two segmental dimensions: F1 & F2 (Hz). Tokens were segmented in Praat (Boersma 2007) using text grids aligned to the start and end of the vocalic section within each syllable. Values were extracted with the help of ProsodyPro (Xu 2013) and FormantPro (Xu and Gao 2018).

2.4. Statistical analysis

Tokens were labelled according to the presence or absence of stress (`stress` = 1 or 0), and according to a combination of their stress value and their position in the disyllabic word (`syllpos` = l0, l1, r0 or r1), where 'l1' refers to a stressed syllable in the left position (i.e. con1) and 'r0' to an unstressed syllable on the right (i.e. flict0). Dynamic articulatory and acoustic data (F0, F1 & F2) were time-normalised, within-speaker z-scored and compared using generalised additive mixed modelling (GAMM) in R (based on recommendations from Wieling 2018). Statistical analyses of articulatory gestures and averaged acoustic data were within-speaker z-scored and carried out using linear mixed-effects modelling with the lme4 package (Douglas Bates, Bolker, and Walker 2015) and post hoc pairwise comparisons were calculated using the EMMEANS package (Lenth et al. 2018).

3. Results

3.1. Consonant gestures

Density plots for gestural plateau duration, stiffness and peak velocity are shown in Figure 2. Separate linear mixed-effects models were fitted for each measurement, with stress as the predictor. Random slopes and intercepts were included for syllable type. Among these three variables, only gesture duration demonstrated a statistically significant association (p = 0.005) - indicating that gesture duration is longer in stressed syllables. Stiffness (p = 0.299) and peak velocity (p = 0.415) did not show a significant relationship with stress. A linear mixed-effects model was fitted with `syllpos` as the independent variable (see Figure 3), and post hoc pairwise comparisons were conducted using estimated marginal means (EMMs) with Tukey adjustment for multiple comparisons. Out of the six combinations, only r0 - r1 was found to be significant (est. -0.564, p = 0.0063).

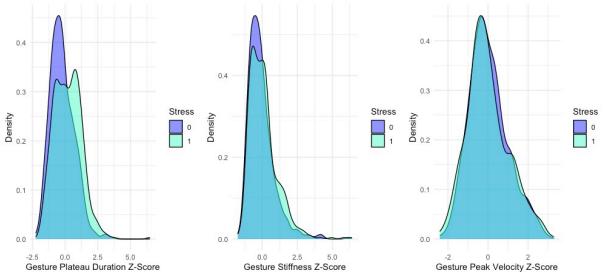


Figure 2: *Density plots of syllable initial consonant measurements; 0 = unstressed, 1 = stressed*

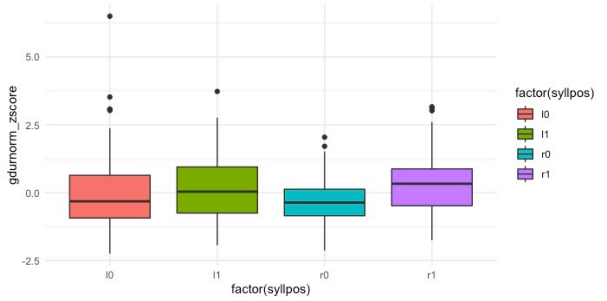


Figure 3: *Gesture plateau duration by syllable position; l0 = left position - unstressed, l1 = left position - stressed, r0 = right position - unstressed, r1 = right position - stressed*

3.2. CV lag

Only the syllable pair 'DI/di' was subjected to CV lag analysis due to the challenges in delineating gestures for other stimuli. These challenges arose from shared articulators between consonants and vowels, along with complex syllable onsets, which made it difficult to clearly identify relevant gestures. Figure 4 presents density plots for stressed and unstressed pair: 'DI/di'. Another linear mixed-effects model was fitted, using the same formula as that used for the consonant gestures, with z-scored, time-normalised CV lag as the dependent variable. The stressed syllable 'DI' had significantly longer CV lag than unstressed 'di' (p = 0.0001).

Figure 4: *Density plots of z-scored time normalised CV lag for 'di/DI'; 0 = unstressed, 1 = stressed*

3.3. Acoustic means

Linear mixed-effects models were fitted for each of the suprasegmental measurements of interest (mean values of in-

Table 1: Vowel GAMM articulatory / F1-F2 results (x = front/back; z = high/low)

	TDz	TDx	TBz	TBx	TTz	TTx	JAWz	JAWx	F1	F2
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	* high	* low
JECT					* inferior		* inferior		* low	* high

tensity, F0 and duration) using the same formula used in the consonant measurement analysis. All acoustic measurements showed significant differences between stressed and unstressed syllables. Stressed syllables were associated with higher values of mean intensity (p = 1.7e-06), mean F0 (p = 0.0001) and mean duration (p = 0.005). As with gesture durations, a linear mixed-effects model was fitted with the factor `syllpos` as the independent variable and post hoc pairwise comparisons were computed. For intensity, all combinations of stressed versus unstressed syllables differed significantly, only l0 - r0, and l1 - r1 showed no significant difference. F0 differed significantly for all combinations other than l0 - r0. The combination of l1 - r1 differed significantly, indicating higher F0 in stressed syllables in the left position of the disyllabic word than those in the right position (p = 0.0055). Duration differed significantly between l0 - l1, l1 - r0 and, as with F0, between l1 - r1 (with stressed syllables on the left being longer. p = 0.0386). Interestingly, the difference in acoustic duration between r0 and r1 did not reach statistical significance, contrasting with the gestural duration analysis, in which r0 - r1 was the only significantly different combination.

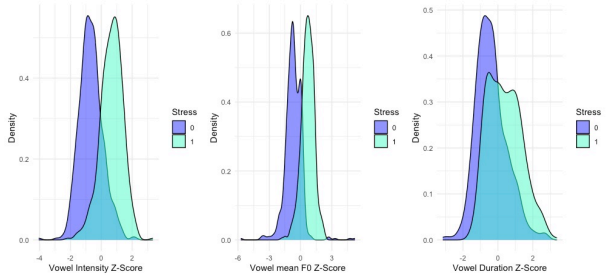


Figure 5: *Density plots of acoustic measurements of vowels; 0 = unstressed, 1 = stressed*

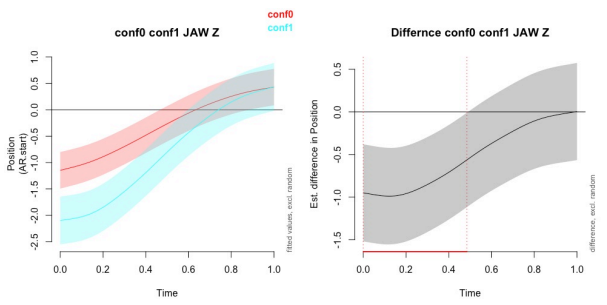


Figure 6: *GAMM smooth and difference plots of JAWz (vertical movement) for 'con'/'CON'*

3.4. Dynamic articulatory & acoustic analysis

3.4.1. Articulator and formant trajectories

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 (see Figure 6). 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. Similarly, formant values are labelled 'high' or 'low' to denote the stressed syllable's relative value during the window of significant difference.

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. Significantly different formant trajectories were found only for 'PRO' and 'JECT'. The other four syllables showed significant differences in articulatory trajectories despite showing no acoustic difference in terms of the first and second formants.

3.4.2. F0 trajectory

A GAMM analysis of F0 trajectories over time across the four levels of `syllpos` revealed that stressed syllables are not only produced with higher F0 but also appear to be produced with distinct tonal contours as a function of the interaction between stress and the position within the disyllabic word. As shown in Figure 7, stressed syllables in the left position are produced with a steady high tone, whereas stressed syllables on the right drop abruptly from their highest point, roughly approximating Mandarin Chinese's 'first' and 'fourth' tones, respectively. Unstressed syllables on the right descended into creaky voice phonation for most of the speakers, whereas those on the left we produced with a more level tone in the middle of the speakers' pitch range.

4. Discussion and conclusion

The results of the acoustic analysis showed that, suprasegmentally, stressed syllables were found to be positively correlated with F0, duration and intensity. Segmentally, significant differences were observed in the vowel formant trajectories between 'PRO/pro' and JECT/ject', whereas no such differences were found between 'CON/con' and 'GEST/gest', despite these syllables sharing a phonemic vowel and occupying identical positions within the word. That stress is a associated with F1

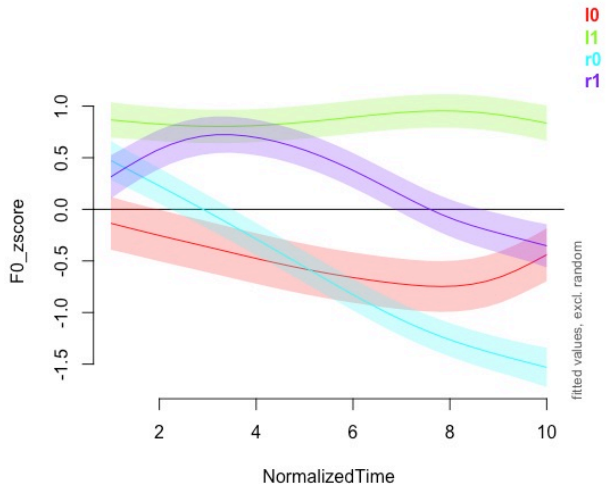


Figure 7: *F0 contours by syllable position and stress; l0 = left position - unstressed, l1 = left position - stressed, r0 = right position - unstressed, r1 = right position - stressed*

& F2 difference in 'PRO' but not 'CON', could be to do with the effect of the articulation of the rhotic in 'PRO'. Regarding 'JECT' and 'GEST', data visualisation revealed a pattern of centralisation in 'GEST' similar to that observed in 'JECT'; however, this difference did not reach statistical significance. GAMM analyses of F0 trajectories further uncovered distinct F0 contours arising from the interaction between stress and syllable position, details that would have been overlooked in a study focusing on measurements from static points.

The results of the articulatory analysis suggest that in Taiwanese Mandarin-accented English, stressed vowels correlate with greater jaw displacement and exhibit significant 'hyper-articulation' of the lingual articulators, even when the vowels are segmentally identical in terms of their first and second formants. Furthermore, consonant gesture plateau duration and CV lag were found to be significantly longer in stressed syllables. Interestingly, Kim's (2021) study suggests that the stressed syllables do not involve substantial supra-glottal hyper-articulation in L2 English by speakers of Standard Chinese. 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.

5. Acknowledgements

This study was supported by a research grant (MOST 110-2410-H-007-025), awarded to the second author, to which we are grateful.

6. References

Boersma, Paul (2007). "Praat: doing phonetics by computer". In: *http://www. praat. org/*.

Delattre, Pierre (1969). "An acoustic and articulatory study of vowel reduction in four languages". In.

Douglas Bates, MM, Ben Bolker, and Steve Walker (2015). "Fitting linear mixed-effects models using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48.

Fry, Dennis B (1955). "Duration and intensity as physical correlates of linguistic stress". In: *The Journal of the Acoustical Society of America* 27.4, pp. 765–768.

Katsika, Argyro and Karen Tsai (2021). "The supralaryngeal articulation of stress and accent in Greek". In: *Journal of phonetics* 88, p. 101085.

Kim, Boram (2021). "Lexical Stress Realization in Mandarin Second Language Learners of English: An Acoustic and Articulatory Study". In.

Lenth, Russell V., Henrik Singmann, Jonathon Love, Paul Buerkner, and Maximilian Herve (2018). *Package "Emmeans"*. R package version 4.0-3. URL: `http://cran.r-project.org/package=emmeans`.

Lieberman, Philip (1960). "Some acoustic correlates of word stress in American English". In: *The Journal of the Acoustical Society of America* 32.4, pp. 451–454.

Pierrehumbert, Janet Breckenridge (1980). "The phonology and phonetics of English intonation". PhD thesis. Massachusetts Institute of Technology.

Roon, Kevin D, Philip Hoole, Chakir Zeroual, Shihao Du, and Adamantios I Gafos (2021). "Stiffness and articulatory overlap in Moroccan Arabic consonant clusters". In: *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 12.1, pp. 1–23.

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

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

Xu, Yi (2013). "ProsodyPro—A tool for large-scale systematic prosody analysis". In: Laboratoire Parole et Langage, France.

Xu, Yi and Hong Gao (2018). "FormantPro as a tool for speech analysis and segmentation". In: *Revista de Estudos da Linguagem* 26.4, pp. 1435–1454.

Predicting Articulatory Landmarks with Critically-Damped Oscillators and General Tau Theory

Christopher Geissler[1], Jyothiraditya Nellakra[1]

[1]*Carleton College, USA*

cgeissler@carleton.edu, nellakraj@carleton.edu

Abstract

Dynamical systems are useful for bridging discrete and continuous aspects of speech. In this paper, we compare the ability of two models, critically-damped oscillators and General Tau Theory, to predict gestural landmarks.

Predictions of the two models were compared with each other and with original kinematic data. The data consisted of electromagnetic articulography recordings of Tibetan collected as part of Geissler (2021). In addition to the landmarks-based approach, this study also uses a a language with a different phonological typology.

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 farther from the target. These results highlight the differences in, and invite further comparison on, the trajectory shape generated by the two models.

Keywords: articulation, articulatory phonology, gestures

1. 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.

2. Methods

Predictions of the two models–critically-damped oscillators and General Tau Theory–were compared with each other and with original kinematic data. Data and code are available on OSF: `https://osf.io/x34sa/`

2.1. Kinematic data

The data consisted of electromagnetic articulography recordings collected as part of Geissler (2021). Six native speakers (four female) of Tibetan living in and around New York City participated in the experiment. All speakers were multilingual, and all speakers were raised in Tibetan diaspora communities in India and Nepal.

Stimuli consistent of Tibetan words elicited in a carrier sentence, presented on a screen in the Tibetan orthography. Target syllables were word-initial and consistent of /m/ followed by the vack vowels /u o a/. The target words were preceded by the vowel /i/ in the carrier sentence in order to facilitate identification of vowel retraction. Target syllables were balanced to include both high and low tone, presence/absence of a coda consonant, and occured either in one-syllable words or as the first syllable in a two-syllable word.

EMA sensors were placed on the upper and lower lips, lower incisor, tongue tip, dorsum, and blade. Consonant gestures were identified as the closing of the lips, and the vowel gesture was identified as the retraction of the tongue dorsum. Gestural landmarks, depicted in Figure 1, were calculated in *Mview* (Tiede 2005), and the position, velocity, and timestamp of each landmark was recorded. In the closure phase, **Gestural Onset** and **Nuclear Onset** were defined as the points at which 20% of peak velocity were achieved in acceleration and deceleration toward the target. Likewise, **Nuclear Offset** and **Gestural Offset** were defined as points with 20% of peak velocity in movement away from the target. These timestamps, along with the point of **Maximum Constriction**, were the focus of analysis.

2.2. Simulations

Parameters for each model were set using certain landmarks, then used to predict the spatio-temporal coordinates at other landmarks. Both models took as inputs the displacement (change in position) of a gesture; the critically-damped oscillator model used the peak velocity and the point at which this was achieved (PVEL and PVEL2), while the General Tau model also used the duration of the movement.

Both models could then calculate the position at any given

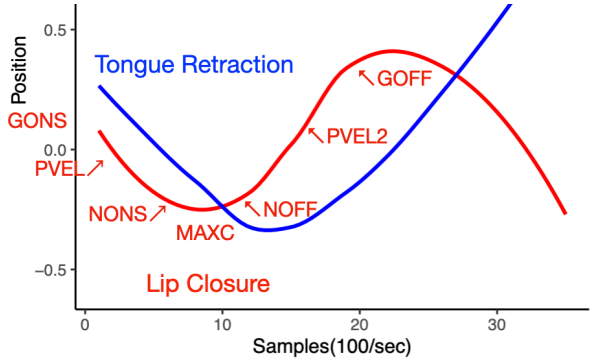


Figure 1: Gesutral landmarks in the lip closure gesture of [ma]. GONS = gesture onset; PVEL = peak velocity of closure; NONS = nuclear onset; MAXC = maximum constriction; NOFF = nuclear offset; PVEL2 = peak velocity of release; GOFF = gesture offset

time duration the gesture, and were used to identify the position and timestamps for the gesture-internal landmarks PVEL, NONS, and PVEL2.

2.2.1. *Critically-damped oscillator model*

In the critically-damped oscillator model, it is possible to calculate the position from a timestamp (or vice versa) using two parameters: the displacement and the natural frequency of the oscillator. The displacement was calculated as the distance from gestural onset to maximum constriction for the closure phase, and the distance from nuclear offset to gestural offset for the release phase. The natural frequency, ω_0, can be calculated from the position and velocity of the system at the point of peak velocity. (1) shows the general equation for a mass-spring system, which can be restated as (2) for a critically-damped oscillator.

$$m\ddot{x} + b\dot{x} + kx = 0, \tag{1}$$

$$\ddot{x} + 2\omega_0\dot{x} + \omega_0^2 x = 0. \tag{2}$$

At the point of peak velocity, this simplifies to (3), since the acceleration $\ddot{x} = 0$. Note that, since the oscillator returns to an equilibrium point $x = 0$, the velocity will always have a sign opposite to the displacement, which ensures that the value of ω_0 must be positive.

$$\omega_0^2 x_{\text{PVEL}} = -2\omega_0 \dot{x}_{\text{PVEL}} \implies \omega_0 = -2\left(\frac{\dot{x}_{\text{PVEL}}}{x_{\text{PVEL}}}\right) \tag{3}$$

Thus, by knowing the displacement x_0, peak velocity, and position at peak velocity, the position x can be calculated as a function of time t using (4):

$$x(t) = x_0\left(e^{-\omega_0 t} + \omega_0 t e^{-\omega_0 t}\right), \tag{4}$$

2.2.2. *General Tau model*

For the Tau model, we used the following equation from Elie, Lee, and Turk (2023), which is derived from Lee (1998). This gives the position of an articulator at a given time t from its starting position x_0 and T, the time at which the target ($x = 0$) is to be achieved.

The only additional parameter is κ, which is analogous to stiffness in that it determines the shape of the velocity profile. $\kappa = 0.4$ was used, following the observation by Elie, Lee, and Turk (2023) that this value held across speakers and articulators; this is also the value at which velocity profiles are symmetrical.

$$x(t) = x_0\left(1 - \frac{t^2}{T^2}\right)^{\frac{1}{\kappa}} \tag{5}$$

The Tau model thus requires the displacement and duration of a movement in order to predict the points in between. The displacement was the same as for the critically-damped oscillator model: the distance from gestural onset to maximum constriction for the closure phase, and the distance from nuclear offset to gestural offset for the release phase. The timestamps of these points were used as the durations.

3. Results

A comparison of predicted with actual data is presented in the figures below: **Figure 2** for position data, and **Figure 3** for temporal data. 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.

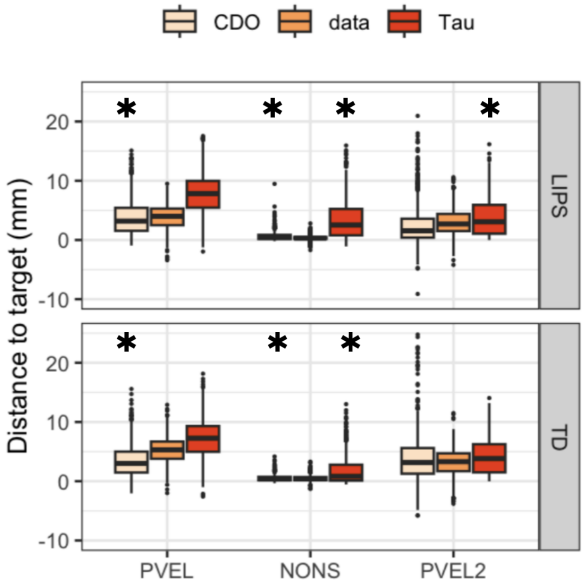


Figure 2: *Predicted position for landmarks in Tibetan /mV/ sequences. Asterisks indicate significant differences between predicted and observed data. CDO = critically-damped oscillator; data = kinematically-defined landmarks; Tau = General Tau model. PVEL/PVEL2 = point of peak velocity toward/away from target; NONS = (gestural) nucleus onset*

We performed a linear mixed-effects analysis on the relationship between these data and their source (kinematic data, oscillator model, Tau model) using the *lme4* package in R.

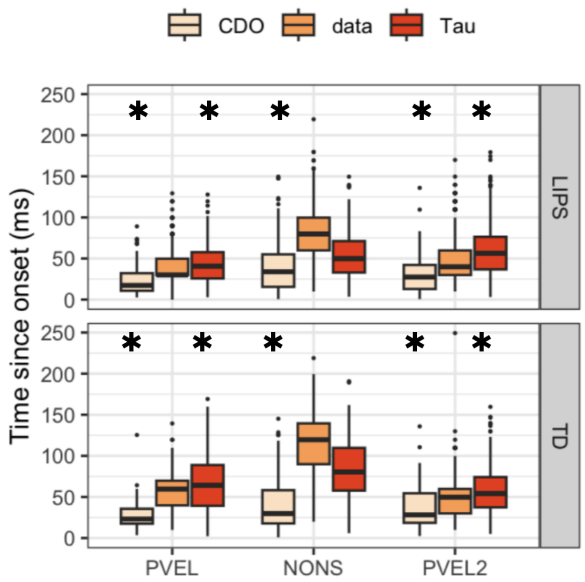


Figure 3: *Predicted time for landmarks in Tibetan /mV/ sequences. Abbreviations as in 2*

We fit two models: one for the position data, and one for the time data. For each, we entered as fixed effects the landmark (PVEL, NONS, PVEL2), articulator (lips or tongue dorsum), and source, as well as random effects of speaker and word. These models were compared to another pair of models that also included an interaction between landmark and source. **Table 1** reports this model comparison, which supports the model that includes an interaction.

Table 1: *Comparison of baseline and interaction models, showing improved fit with interaction.*

Position Model	AIC	BIC	logLik
baseline	184609	184687	-92295
interaction	182684	182797	-91329
Time Model	AIC	BIC	logLik
baseline	278535	278609	-139258
interaction	277350	277457	-138662

We conducted a post-hoc analysis using the *emmeans* package to identify pairwise differences between levels of the models. Specifically, we noted where there were significant differences between oscillator- or Tau-predicted data and the observed kinematic data. These are indicated in Figs. 2 and 3.

Predictions of the oscillator model were significantly different from kinematic data in 10 of 12 cases, while the predictions of the Tau model were significantly different in 7 cases. Interestingly, the Tau model achieved closer values than the oscillator model on the peak-velocity landmarks (PVEL and PVEL2) despite the fact that the oscillator model used these points as inputs.

The direction of the divergence between models is also noteworthy. In the spatial domain, the oscillator model tended to predict that landmarks would occur slightly closer to the target than was identified in the kinematics, while the Tau model predicted landmarks occurring slightly farther from the target. In the temporal domain, the oscillator model predicted landmarks occurring earlier than in the kinematics, while the Tau-predicted landmarks occurred around the same time as, or after, their kinematic equivalents.

4. Discussion

This study compared the ability of two models to predict the spatial and temporal points at which kinematically-defined gestural landmarks would occur. Both the critically-damped oscillator model and the General Tau model predicted landmarks with a fair degree of accuracy, but with some systematic differences. Oscillator-predicted landmarks fell sooner and closer to the target, while the opposite was the case for the Tau-predicted landmarks.

These results highlight the differences in the shapes of the trajectories generated by each model. Critically-damped oscillators move rapidly, then slow to asymptotically approach the target; Tau-derived trajectories unfold gradually (and, when $\kappa = 0.4$, symmetrically), and reach the target at a known point in space and time. We encourage further research on the models to address not only overall fit to data, but also how the details of particular shapes.

The use of data from a less-commonly studied language, Tibetan, is an important part of creating models that more accurately capture the diversity of human speech. It is noteworthy that the value of κ obtained from English speech by Elie, Lee, and Turk (2023) worked reasonably well for the Tibetan data. Further study is needed on the ways κ might vary across languages, speakers, natural classes, articulators, and contexts, parallel to similar work on stiffness in Task Dynamics.

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.

5. Acknowledgements

This research was supported by a Student Research Partnership grant from the Humanities Center at Carleton College. Original data collection was supported by Yale University, and conducted with the help of Jason Shaw and Muye Zhang.

6. 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).

Lee, David N. (Sept. 1998). "Guiding Movement by Coupling Taus". In: *Ecological Psychology* 10.3/4, p. 221. DOI: 10.1080/10407413.1998.9652683.

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.*

Allophones of Korean /l/: a classification using EMA

Kye Shibata[1], Feng-fan Hsieh[1], Yueh-chin Chang[1]

[1]*National Tsing Hua University, Taiwan*

kye.shibata@gmail.com, ffhsieh@mx.nthu.edu.tw, ycchang@mx.nthu.edu.tw

Abstract

In this paper, an electromagnetic articulography (EMA) study is conducted on four standard Korean speakers to examine their articulation of Korean /l/. Their articulation of /l/ is compared to another sonorant, /n/, in the context of the vowels /a/, /i/, and /u/. Lateralization is quantified by calculating the lateralization angles of both the left- and right-side parasagittal sensors, and retroflexion is examined by observing the angle of elevation seen in the tongue-tip sensor. Considerable intra- and inter-speaker variance is observed in the articulation of /l/, with a general tendency for /l/ to be asymmetrically lateralized in coda position, suggesting a lateral approximant articulation. Lateralization appears to be optional in onset position, but the tongue tip is generally raised, indicating a flap, retroflex lateral, or lateral flap articulation. Coda /l/ is also observed to have a tendency to have a raised tongue tip in the context of /a/, suggesting it is a retroflex lateral approximant in this environment.

Keywords: electromagnetic articulography, Korean, lateral, lateralization, flap

1. 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, Charles, and Lulich 2019; Lee, Goldstein, and Narayanan 2015), no study to our knowledge has utilized EMA.

Sohn (1999) and Shin, Kiaer, and Cha (2013) describe two allophones of /l/: an alveolar lateral approximant [l], appearing word-finally or word-medially as a geminate, and an alveolar tap [ɾ], appearing word-initially and word-medially as a singleton. Crosby and Dalola (2021) conducts a formant study of Korean /l/ and finds that utterance-final instances of /l/ were often retroflexed–though there is considerable inter-speaker variation. Hwang, Charles, and Lulich (2019) conducts a 3D ultrasound study of word-final Korean /l/, also concluding that there is significant inter-speaker variation in both place or articulation and size of occlusion, finding apico-dental, lamino-postalveolar, lamino-alveolar, and retroflex articulations, with considerable asymmetry being observed for larger occlusions.

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.

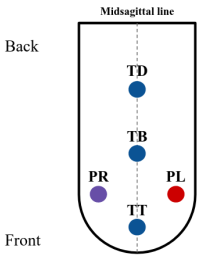


Figure 1: *"Southern Cross" configuration of sensors*

2. Methods

Data was collected from four standard Korean speakers in their 20s (two female, two male) using the Carstens' Articulograph AG501. Five sensors (TT = tongue tip, TB = tongue blade, TD = tongue dorsum, PR = speaker-right parasagittal, PL speaker-left parasagittal) were placed on the speaker's tongue, following the "Southern Cross" configuration described in Strycharczuk, Derrick, and Shaw (2020) and pictured in **Figure 1**. The speakers were asked to read a mix of real Korean words and nonce words written in hangul in the carrier sentence "*Jigeum* _____ *bwa*", meaning 'Now look at _____'. We included words that had both /l/ and /n/ in onset and coda positions, in the context of the vowels /a/, /i/, and /u/, yielding the target sequences /la/, /na/, /li/, /ni/, /lu/, /nu/, /al/, /an/, /il/, /in/, /ul/, and /un/. Stimuli containing each sequence were repeated 7 times by each speaker, giving 84 tokens per speaker.[1]

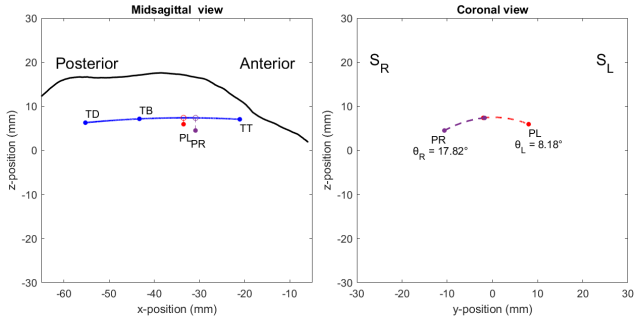
The data was processed using MVIEW (Tiede 2005), as well as custom scripts written in MATLAB. Examples of sensor positions measured during the experiment are shown in **Figure 2**.

To quantify lateralization, we utilized the lateralization angle method described in Huang et al. (2023), where the parasagittal sensor positions are compared with the midsagittal line of the tongue surface (using a spline fit through the three midsagittal sensors) to get the angle at which each parasagittal sensor is lowered. The lateralization angle θ_L is calculated as:

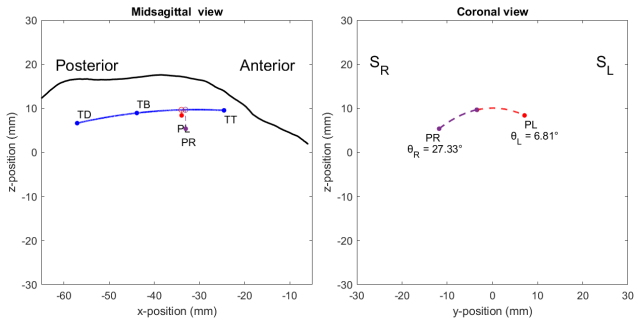
$$\theta_L = \arctan\left(\frac{\alpha d_{xz}}{d_y}\right), \tag{1}$$

where θ_L is the lateralization angle, d_{xz} is the Euclidean distance between the parasagittal sensor and its closest point on the midsagittal line (the midsagittal intercept) on the x-z plane, d_y is the y-dimensional distance between the parasagittal sensor and its corresponding midsagittal intercept, and α is a coefficient which is positive if the parasagittal sensor is lower than

[1]For Speaker 1, only 5 repetitions were recorded, for a total of 60 tokens.



(a) *Onset in /la/*



(b) *Coda in /al/*

Figure 2: *Examples illustrating the onset of /la/ shortly before release, and the coda for /al/ shortly before the end of the rime for Speaker 1. The black line is a trace of the speaker's palate. The speaker is facing right.*

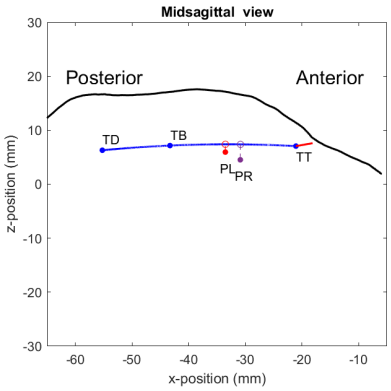
its corresponding midsagittal intercept in the z-dimension, and negative if higher.

The lateralization angle was calculated at a specific time across all trials for each speaker in each vowel context. For onset /l/, measurements were taken at 90% of the entire duration of the onset, to examine the tongue posture shortly before release. For coda /l/, measurements were taken at 80% of the duration of the entire syllable, where the /l/ coda was found to be steady and not yet affected by the following syllable. The same was done for /n/.

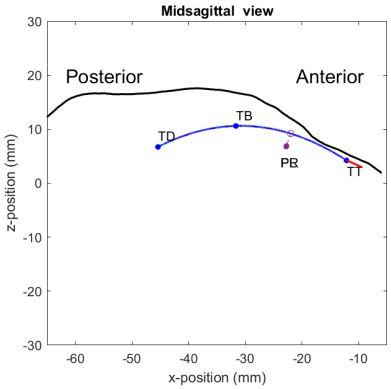
To examine whether retroflexion occurred, the angular information provided by the AG501 for each sensor was taken into consideration. Specifically, the elevation angle for the TT sensor (placed approximately 1 cm behind the tongue tip) was examined to determine whether there was a raised tongue tip. Two examples are illustrated in **Figure 3**, where a red line extended from the TT sensor illustrates the angle of elevation. The elevation angle for onset /l/ was taken at 90% of the entire duration of the onset, and for coda /l/ it was taken at 80% of the entire syllable, as above.

3. Results

The lateralization angles calculated from the two parasagittal sensors for /l/ and /n/ in both onset and coda positions in each vowel context are given for the four speakers (S1 = Speaker 1, S2 = Speaker 2, S3 = Speaker 3, S4 = Speaker 4) in **Figure 4**. The four speakers showed great variability in their articulation of /l/ in both onset and coda position.



(a) */la/*



(b) */na/*

Figure 3: *Examples illustrating the TT elevation angle for /la/ and /na/ in Speaker 1, just before the release of the onset. Red line indicates the angle of elevation.*

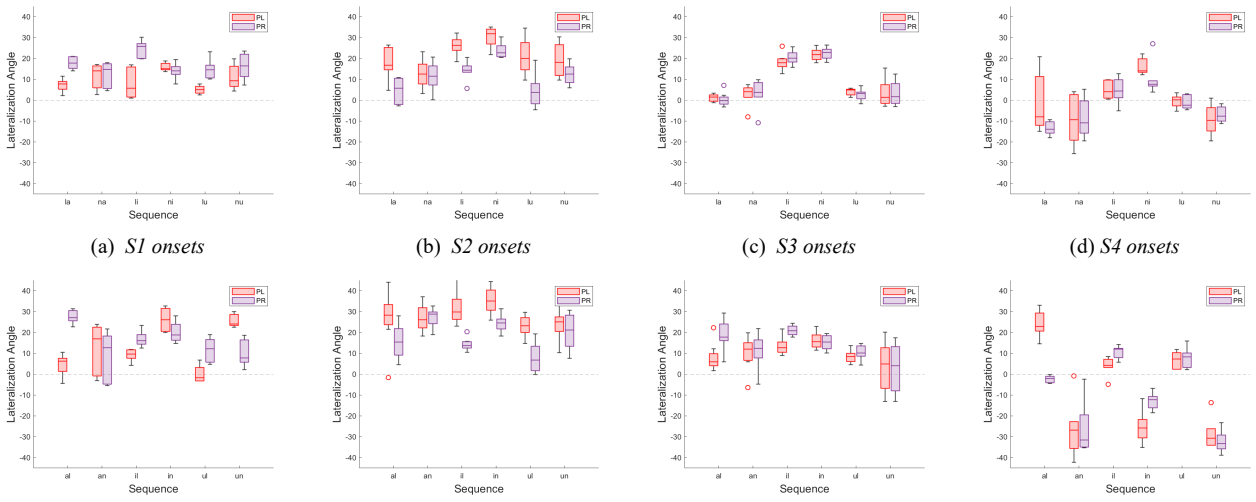
3.1. Onset /l/

For the onset, S1 had consistently positive lateralization angles, indicating a lowering of that side of the tongue. There was also a stark asymmetry in the left- and right-side lateralization for /l/ that was not observed for /n/. In all vowel contexts, it was found that the PR was consistently lower than the PL sensor for /l/, while for /n/ the lateralization angles for both the PL and PR sensors were similar.

S2 had a similar pattern, with asymmetric lateralization being observed for /l/ and not /n/ (though /n/ had some asymmetry in the context of /i/, it was less than what was found for /l/). However, unlike S1 who had right side lowering, S2 had a preference to lower his left side. Lateralization angles were generally positive in all contexts for both /l/ and /n/.

S3 had almost no difference in right and left side lateralization for both /l/ and /n/, and differences between /l/ and /n/ were minimal. Almost all measurements had a positive lateralization angle, though some anti-lateralization (raising of the side of the tongue) can be observed for the /la/ and /na/ sequences.

S4 showed the greatest degree of variance. The sequence /la/ had the greatest variance, with left-side lateralization angle values differing wildly between a lowered tongue and a raised tongue depending on the token. Other than for /la/, the only consistent asymmetry between the two sides was in /ni/, where the left side had slightly greater lowering. S2 was observed to have anti-lateralization for both sides when followed by /a/ and



(a) *S1 onsets* (b) *S2 onsets* (c) *S3 onsets* (d) *S4 onsets*

(e) *S1 codas* (f) *S2 codas* (g) *S3 codas* (h) *S4 codas*

Figure 4: *Lateralization angle values for /l/ and /n/ in onset and coda position. Red box plots are the left side values, while the purple box plots are the right side.*

/u/, for both /l/ and /n/.

3.2. Coda /l/

For the codas, S1 showed consistent right-side lowering for coda /l/, as observed for onset /l/. Coda /n/ showed less obvious asymmetry for /an/ (which had considerable variance) and /in/, but for /un/ it was found that the left side had greater lowering, going against the right-side preference observed in the articulation of /l/.

S2 had a similar pattern to S1, with asymmetry being found for coda /l/ but less so for coda /n/. Once again, S2 had a preference for left-side lowering, as observed in the onset /l/ examples. The /n/ codas had little to no asymmetry in /a/ and /u/ contexts, but in the context of /i/ there was a greater lowering of the left side, though less than what was observed for the /l/ coda.

S3 also had a more notable asymmetry in lateralization for /l/ than for /n/. Specifically, the /l/ coda was associated with greater lateralization of the right side of the tongue, while /n/ codas had roughly equivalent lateralization on either side. Much like for the onsets, lateralization was generally positive, with the exception of /un/ and /an/ which showed some anti-lateralization.

S4 showed some asymmetry in lateralization, and showed a strong tendency for positive lateralization for /l/ codas, while having anti-lateralization for /n/ codas. Asymmetry was most notable for /al/, in which the left side of the tongue was lowered much more than the right side, which was slightly raised. However, unlike with the consistent right-side lowering seen in S1, S2 had some degree of randomness–for /il/ the right side was lowered slightly more than the left, and for /ul/ there were similar values for either side recorded. As noted with /ni/, /in/ had asymmetry as well, though with the right side being less raised than the left side in this case.

3.3. Retroflexion

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 and S4 were disregarded as the angular information was not usable due to issues with the sensor placement). In general, both S1 and S3 had a more posterior place of articulation for /l/ compared to /n/, in both onset and coda positions. The values for the elevation angle of the TT sensor are shown in **Figure 5**.

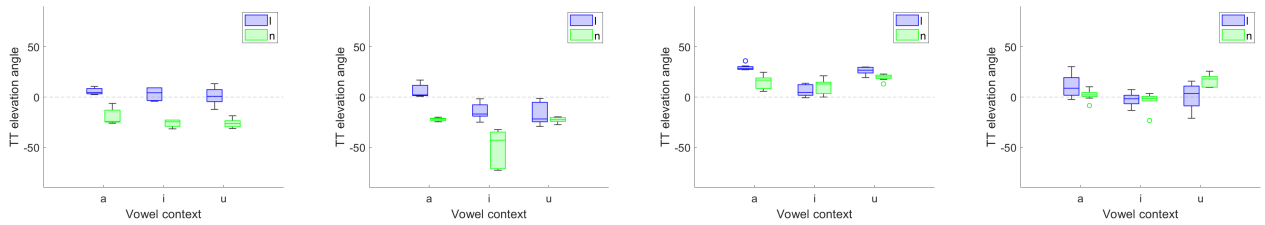
S1 had a consistent pattern for the onsets, with /n/ exhibiting a negative elevation angle for all vowel contexts, indicating a downward oriented TT. On the other hand, /l/ had a slightly positive elevation angle consistently, indicating a slightly raised or retroflexed tongue tip. Where /l/ was followed by /a/, the angle was consistently positive, while when followed by /i/ or /u/ there was either slight retroflexion or a slightly downward orientation. In coda position, /l/ had consistent retroflexion in the /a/ context, but a downward orientation in the context of /i/ and /u/. Once again, /n/ was always articulated with a downward orientation of the TT. In all contexts, /l/ tended to have a more upward or flat TT orientation than /n/.

In onset position, S3 generally had positive values for the TT elevation. The speaker had a more raised TT for /l/ than /n/ when followed by either /a/ or /u/, especially in the context of /a/. For /i/, both /l/ and /n/ showed less elevation of the TT. In coda position, S1 had less retroflexion in general, but /l/ in the context of /a/ exhibited retroflexion sometimes. In the context of /i/, there was little retroflexion, or sometimes a slight downward angle of the TT. In the context of /u/, /l/ had some variance, but varied between somewhat retroflexed and somewhat downward TT orientations. Meanwhile, /n/ exhibited consistent retroflexion in the context of /u/.

4. Discussion and conclusion

The results indicate that lateralization for /l/ and /n/ varies based on the speaker.

S1 consistently had a lowered right side of the tongue when articulating /l/, in both onset and coda position. This indicates that /l/ is articulated as a lateral in all contexts. In onset position, and in coda position following /a/, there is also some degree of tongue-tip raising, suggesting these could be classified



(a) *S1 TT elevation angle, onsets* (b) *S1 TT elevation angle, codas* (c) *S3 TT elevation angle, onsets* (d) *S3 TT elevation angle, codas*

Figure 5: *Angle of elevation for the TT sensor. Blue boxes indicate the values for /l/, green boxes indicate the values for /n/.*

as instances of a retroflex lateral, or perhaps a retroflex lateral flap. S2 had a very similar pattern, though with a preference for lowering the left side of the tongue. Once again, the onset /l/ should be classified a lateral, though whether or not retroflexion occurred is unclear due to issues with the angular data for S2.

S3 had clear, asymmetric lowering of the right side when /l/ was in coda position, indicating clear lateralization. In onset position, there was no notable asymmetry, and in fact /l/ and /n/ had similar values for lateralization in the same vowel context for both sides. This seems to be in line with previous classifications of Korean /l/, where it is a non-lateral flap in onset position, but a lateral in coda position. In addition, there was retroflexion observed in coda position, but only in the sequence /al/, indicating that in this context /l/ was articulated as [[].

S4 had a much more varied lateralization pattern, with some contexts having left-preferred lateralization, but not others. S4 also had a noticeable tendency for anti-lateralization, something that was not seen in the other speakers. These differences are likely due to a combination of factors, such as the shape of the speaker's palate. It is difficult to classify what was observed in the data for S4, except the fact that both /la/ and /al/ sequences have /l/ articulated with lateralization, and that they exhibited a preference for lowering the left side, as seen in S2.

Direct comparison of the data in statistical analyses is difficult, as all four speakers exhibit starkly different patterns in their articulations. In the future, data from more speakers could allow for a rough classification of speakers into at least two groups, those who prefer right-side lateralization and those who prefer left-side lateralization.

Coda /l/ in all four speakers had a tendency to have greater asymmetry between the two sides than /n/, 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] in most cases, as indicated in the literature. 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 [[].

In onset position, there was significant variation across speakers. While S1 and S3 indicated TT raising for all instances of onset /l/, asymmetric lateralization was only observed in S1 and S2, suggesting that some speakers articulate it as an alveolar flap (S3), while others articulate it as a retroflex lateral or lateral flap (S1 and S2). S4 had inconsistent lateralization in onset position, and had high variance between trials, suggesting that some speakers may be less consistent in their articulation of onset /l/. It is unclear if such variance is common based on the limited data gathered for this study.

It is difficult to provide a generalization that covers all speakers due to the limited amount of data we have available, and due to the variance observed. However, our findings suggest that Korean /l/ has more allophonic variation than indicated in the literature. Specifically, lateralization appears optional in onset position, with some speakers having a clear, asymmetric lateralization and others having almost none. In coda position there is more consistent lateralization, though there appears to also be some degree of retroflexion in the context of /a/.

5. References

Crosby, Drew and Amanda Dalola (2021). "Phonetic variation in the Korean liquid phoneme". *Proceedings of the Linguistic Society of America* 6(1): 701–712.

Huang, Jing, Kye Shibata, Feng-fan Hsieh, Yueh-chin Chang, and Mark Tiede (2023). "The L~N merger in Southwestern Mandarin: an articulatory study". *Proceedings of the 20th International Congress of Phonetic Sciences*. Prague, Czech Republic.

Hwang, Young, Sherman Charles, and Steven M. Lulich (2019). "Articulatory characteristics and variation of Korean laterals". *Phonetics and speech sciences* 11(1): 19–27.

Lee, Yoon-Jeong, Louis Goldstein, and Shrikanth S Narayanan (2015). "Systematic variation in the articulation of the Korean liquid across prosodic positions". *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, United Kingdom.

Shin, Jiyoung, Jieun Kiaer, and Jaeeun Cha (2013). *The Sounds of Korean*. Cambridge: Cambridge University Press.

Sohn, Ho-Min (1999). *The Korean Language*. Cambridge Language Surveys. Cambridge: Cambridge University Press.

Strycharczuk, Patrycja, Donald Derrick, and Jason Shaw (2020). "Locating de-lateralization in the pathway of sound changes affecting coda /l/". *Laboratory Phonology* 11(1): 21.

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

Ying, Jia, Jason A Shaw, Christopher Carignan, Michael Proctor, Donald Derrick, and Catherine T Best (2021). "Evidence for active control of tongue lateralization in Australian English /l/". *Journal of Phonetics* 86: 101039.

Discovering dynamical models of speech using physics-informed machine learning

Sam Kirkham

Lancaster University, UK
s.kirkham@lancaster.ac.uk

Abstract

Spoken language is characterised by a high-dimensional and highly variable set of physical movements that unfold over time. What are the fundamental dynamical principles that underlie this signal? In this study, we demonstrate the use of physics-informed machine learning (sparse symbolic regression) for discovering new dynamical models of speech articulation. We first demonstrate the model discovery procedure on simulated data and show that the algorithm is able to discover the original model with near-perfect accuracy, even when the data contain extensive variation in duration, initial conditions and target positions, as well as in the presence of added noise. We then demonstrate a proof-of-concept applying the same technique to empirical data, which reveals a small set of candidate dynamical models with increasing levels of complexity and accuracy.

Keywords: speech production, sparse symbolic regression, articulatory phonology, task dynamics, articulatory data

1. Introduction

A fundamental aim in the study of language is the discovery of abstract invariants that underlie the variability observed in performance. For example, speech production involves a set of low-dimensional combinatorial units that are physically realised as a set of variable and high-dimensional motions. How do we best model the relationship? One solution is proposed by Articulatory Phonology/ Task Dynamics (AP/TD), in which phonetics and phonology are isomorphic, with the fundamental unit being the speech gesture: an abstract goal-driven force directing the vocal tract to a target state (Browman and Goldstein 1992; Tilsen 2016; Iskarous 2017).

Saltzman and Munhall (1989) propose a model of the gesture (hereafter abbreviated as SM89) as a critically damped harmonic oscillator (1), where k is a stiffness coefficient, m is a mass coefficient, and the damping coefficient $b = 2\sqrt{m}k$.

$$m\ddot{x} + b\dot{x} + kx = 0 \tag{1}$$

The SM89 model has long been the core gestural equation underpinning AP/TD, but it fails to capture the quasi-symmetrical velocity profiles and time-to-peak velocities typical of empirical data. Byrd and Saltzman (2003) show this can be solved via ramping functions, making gestural activation time-dependent. Sorensen and Gafos (2016) argue that this is an undesirable solution and that empirically realistic trajectories can be achieved by instead allowing the restoring force to be non-linear via a cubic term dx^3 in (2). This also eliminates the need for time dependence once the gesture is initiated.

$$m\ddot{x} + b\dot{x} + kx - dx^3 = 0 \tag{2}$$

This model reproduces many characteristics of empirical velocity profiles, but there may still be some room for improvement. For instance, Elie, Lee, and Turk (2023) advance a general Tau model that outperforms the SG16 model in fitting empirical data. Beyond conventional models of the gesture, there is also considerable scope for further developing task dynamic models of other domains, such as prosodic time-series (Iskarous, Cole, and Steffman 2024), disordered speech (Parrell et al. 2023), and signed languages. In many cases, we might have a lot of data, but lack sufficient predictions of the underlying dynamics to propose a model, or we may seek alternative models that better fit empirical data. This raises a question: how can we efficiently develop new dynamical models of speech?

We solve the problem of model discovery by leveraging recent developments in dynamical systems and machine learning that allow us to learn symbolic equations directly from data (Schmidt and Lipson 2009; Brunton, Proctor, and Kutz 2016). In such cases, we want to find a small number of model terms that expose the underlying dynamics, as opposed to a neural network that may have a very large number of parameters. Underpinning this is symbolic regression, whereby a function f can be approximated from $X, \dot{X}$ – which represent time-varying states $x(t), \dot{x}(t)$ – as a combination of non-linear functions:

$$\dot{X} = \Theta(X)\Xi \tag{3}$$

where $\Theta(X)$ is a library of non-linear functions

$$\Theta(X) = [1 X X^2 X^3 \dots sin X cos X] \tag{4}$$

and Ξ is a vector of coefficients corresponding to the functions in $\Theta(X)$.

$$\Xi = [\xi_1 \xi_2 \xi_3 \dots \xi n] \tag{5}$$

Without any constraints, the above model is likely to produce many non-zero coefficients in Ξ that do not contribute much to the underlying system, adding model complexity and increasing the risk of overfitting. In order to promote sparsity in Ξ, *sparse* symbolic regression optimises for a sparse vector of coefficients for each function in $\Theta(X)$. An example optimisation is Sequential Thresholded Least-Squares, which solves a least squares solution for Ξ, thresholds any coefficients below a value λ, and repeats this process until an optimally sparse model is determined (Brunton, Proctor, and Kutz 2016).

The sparse symbolic regression method outlined above falls into a general class of SINDy (Sparse Identification of Non-linear Dynamics) models. SINDy models can accurately discover the governing equations of known systems, such as chaotic Lorenz and fluid dynamic equations, as well as discover new models in applications such as astrophysics (Pasquato et al. 2022). For more details see Brunton, Proctor, and Kutz (2016) and Champion et al. (2020).

2. Methods

The first step in model discovery is obtaining one or more time-series that represent the output of the system under study. In our case, this is the position and velocity of the vocal tract articulators. We aim to model a single speech gesture, so each trajectory represents a single gesture, defined as the interval between a pair of successive zero crossings in the velocity signal.

The next step is to select a library of candidate functions. From AP/TD research reviewed above, we know that articulatory signals are often well-approximated by polynomial functions, such that a function $f(x)$ can be approximated as a sum of polynomials of increasing order, as in (6), where a_n is the coefficient of each term (note that a_0 is a constant). In this instance, we do not allow interactions between terms, such as $x\dot{x}^2$, but allowing this would be a trivial addition.

$$f(x) = a_0 + a_1 x + a_2 \dot{x} + a_3 x^2 + a_4 \dot{x}^2 + a_5 x^3 + a_6 \dot{x}^3 + \dots \quad (6)$$

A key aspect of SINDy is that we can incorporate physical constraints on the discovered model, such that a discovered coefficient must have a specific value, or two coefficients must be in a particular ratio. To illustrate, take the equation $\ddot{x} = -b\dot{x} - kx$. In order to discover or numerically solve a second-order differential equation, we split it into a series of first-order equations with the introduction of a new variable y, such that $y = \dot{x}$ and $\dot{y} = -by - kx$. If SINDy finds $y = 1.00\dot{x}$ then we can just substitute this value easily into the second equation. If it finds a more complex equation, however, such as $y = 43.62 - 1.55x + 0.90\dot{x}$, then it would yield a final model of $\ddot{x} = -b(43.62 - 1.55x + 0.90\dot{x}) - kx$.

To avoid this level of complexity, we place a physical constraint on y such that $y \stackrel{!}{=} 1.00\dot{x}$. We later show that relaxing this constraint results in models that better fit the data, but also add significant complexity. We implement constraints using the SR3 (sparse relaxed regularized regression) algorithm (Champion et al. 2020), which aims to minimise (7), where $R(W)$ is a regularisation function that acts as a prior on sparsity promotion and λ weights this constraint. Note that $\lambda = \eta^2/2\nu$, where ν determines the closeness of the match between Ξ and W.

$$\min_{\Xi, W} \frac{1}{2}||\dot{X} - \Theta(X)\Xi||^2 + \lambda R(W) + \frac{1}{2\nu}||\Xi - W||^2 \quad (7)$$

We use weighted ℓ_0 regularisation, with a coefficient threshold of $\eta = 0.1$ and $\nu = 1$. A model is discovered for each trajectory and we perform model ensembling over these individual models to arrive at a final model. We evaluate the accuracy of the model by generating a prediction from the discovered model for each token. We then score the accuracy of the predicted trajectory using R^2 and RMSE metrics.

3. Discovering models from simulated data

3.1. Generating simulated data

In order to test the ability of SINDy to discover models from data, we generated a simulated data set with a number of parameters varied across a set of trajectories. Specifically, we simulated data across combinations of duration = {0.05, 0.10, 0.15, 0.20} seconds, initial position = {0.0, 0.1, ..., 1.0}, target = {0.0, 0.1, ..., 1.0} and noise = {normal, noise}. In all simulations, k = 2000 and $b = 2\sqrt{k}$. The noise condition corresponds to the addition of random Gaussian noise between [0, 1], scaled by a factor of 0.01, to each position and velocity sample from the simulated solution. We removed cases from the above parameter combinations where the target was equal to the initial position, as the trajectory does not move from its initial condition in these instances. These parameters were used as inputs to the SM89 second-order differential equation $\ddot{x} + b\dot{x} + kx = 0$ which was solved numerically using the `scipy.integrate.solve_ivp` function in Python. This resulted in 880 unique simulated trajectories.

3.2. Results

We perform SINDy discovery on the SM89 model using a simple candidate library containing the terms x and $\dot{x}$, which means that the maximal equation is:

$$\ddot{x} = a_0 + a_1 x + a_2 \dot{x} \quad (8)$$

The SINDy models finds equation (9) for all trajectories. Note that SINDy reports the target as kC, but we can substitute $kx - kC$ with $k(x - C)$. As such, we correctly identify the original equation that simulated the data, even in the presence of the variable durations, targets, initial conditions and noise.

$$\ddot{x} = -b\dot{x} - k(x - C) \quad (9)$$

In the no noise condition, parameter estimation is near 100% accuracy, with the difference between real/estimated coefficients at C = 0.01% ($\sigma = 0.01$), k = 0.08% ($\sigma = 0.02$), b = 0.03% ($\sigma = 0.01$). Reconstruction of the simulated trajectories is also highly accurate, with mean $R^2 = 1.00$ ($\sigma = 0.01$) and mean RMSE = 0 ($\sigma = 0.02$). The addition of noise affects parameter estimation to a minor extent, with mean $R^2 = 0.99$ ($\sigma = 0.12$) and mean RMSE = 0.03 ($\sigma = 0.02$). The difference between real and estimated coefficients in the noisy condition is C = 0.57% ($\sigma = 1.20$), k = 3.00% ($\sigma = 3.78$), b = 3.96% ($\sigma = 4.37$). The worst performing noisy trajectory had $R^2 = 0.84$

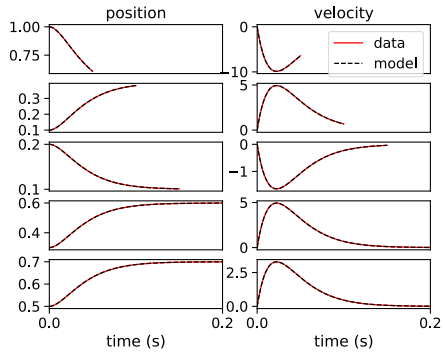


Figure 1: *Simulated trajectories and SINDy predictions for noise-free data. The y-axis varies across each plot to fit the data's range.*

Figure 1 shows 5 randomly sampled trajectories comparing simulated data and discovered model predictions. The model estimates the underlying trajectories with a very high degree of accuracy, even when the data are truncated as in the top two panels. We are unable to show a plot of the noisy data due to space constraints, but reconstruction of the underlying trajectory is also near-perfect in this condition, even in the presence of considerable random noise.

4. Discovering models from empirical data

We now move on to a proof-of-concept example, showing how we can discover parsimonious models from empirical data.

4.1. Data

We use data from the X-Ray Microbeam corpus (Westbury 1994). As a case study, we only analyse data from a single speaker (JW11), as this allows us to explore the initial interpretation of model coefficients, without having to take into account the significant added complexity introduced by between-speaker variation. Specifically, we use a task in which speakers produce a string of repetitions of the syllable /pə pə pə .../. This allows us to examine repetitions of the same gesture, which acts as a valuable test of how sensitive the model discovery procedure is to small variations within one speaker. We see this evaluation as a necessary step prior to applying the method to data with a much greater range of variation. We calculated lip aperture as the Euclidean distance between upper and lower lip sensors, and approximated velocity as the first-derivative of the position values. Gestures were segmented into separate closure and release gestures based on zero-crossings in the velocity signal. In total, we obtained 29 individual gestural trajectories from repetitions of /p/ for this speaker.

4.2. First-order models

We begin by fitting a simple model to the data: a first-order differential equation for $\dot{x}$. Note that here we are only solving for the velocity of the gesture, unlike the SM89 model which solves for acceleration $\ddot{x}$. We predict that a first-order model may be a worse fit for the data than a second-order model, but we begin with a simpler model to assess its baseline accuracy.

Table 4.2 shows a first-order model fitted with different feature libraries of polynomial degrees between one and four. Note that prediction accuracies are for the gesture's position variable only, because SINDy integrates over the velocity to return position. We comment on the model's accuracy in estimating velocity later in this section. A first-degree model performs very poorly with mean R^2 = 0.02, second/third-degree models have mean R^2 = 0.92, and the fourth-degree model has mean R^2 = 0.89. It is clear that the addition of cubic terms has only a negligible effect and the quartic term actively degrades performance, so we now explore this first-order second-degree model further.

degree	R^2 mean	$R^2\sigma$	R^2 min	R^2 max
1	0.02	0.02	0.00	0.07
2	0.92	0.01	0.89	0.94
3	0.92	0.01	0.90	0.94
4	0.89	0.17	0.01	0.94

Table 1: R^2 statistics for first-order models with different polynomial degrees fitted to lip aperture data.

The first-order second-degree model returns a simple quadratic equation:

$$\dot{x} = a - bx + cx^2 \tag{10}$$

There is a linear relationship between a, b, c, such that in these data $a \approx -14b \approx 830c$. As this is a quadratic equation, the quartic term cx^2 determines the width of the velocity peak, the linear term bx controls symmetry around the y-axis, and the constant a determines the y-intercept.

Figure 2 shows randomly sampled lip aperture trajectories and SINDy predictions. We can see very good reconstruction of the position data, but the velocity profiles are less accurate: while the qualitative shape is maintained, the onset/offset are displaced from zero and there are some noticeable mismatches. In summary, a first-order model provides a simple qualitative model that approximates the system, but clearly underperforms in predicting change in velocity. As a result, we anticipate that a second-order model should improve performance.

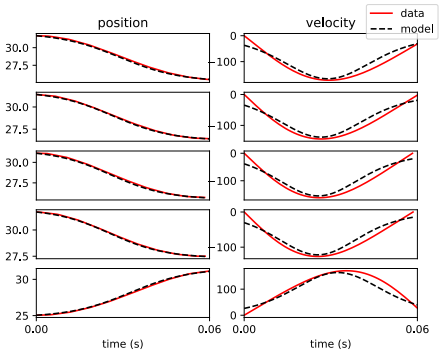


Figure 2: *Lip aperture trajectories and SINDy first-order model predictions, with polynomial terms up to quadratic. The y-axis varies across each plot to fit the data's range.*

4.3. Second-order models

We now fit a second-order model to the data, solving for the system's acceleration $\ddot{x}$. This should allow us to better capture changes in velocity. Note that we impose a physical constraint on the velocity as detailed in Section 2, which simply aims to reduce model complexity and aid interpretability. Table 4.3 shows a second-order model fitted with different feature libraries of polynomial degrees between one and four. The first- and second-degree models have mean R^2 = 0.96, which is slightly better than the higher polynomials. This suggests that a first-degree model can perform well, so we explore this further.

degree	R^2 mean	$R^2\sigma$	R^2 min	R^2 max
1	0.96	0.00	0.95	0.96
2	0.96	0.00	0.95	0.96
3	0.95	0.01	0.92	0.96
4	0.94	0.02	0.90	0.96

Table 2: R^2 statistics for second-order models with different polynomial degrees fitted to lip aperture data.

The second-order first-degree model returns (11), which is equivalent to the Saltzman and Munhall (1989) model.

$$\ddot{x} = -b\dot{x} - k(x - C) \tag{11}$$

Figure 3 shows the same 5 lip aperture trajectories as in Figure 2, with SINDy predictions from the second-order model. The discovered model fits better than the first-order model, but with some inaccuraries towards the end of the velocity trajectory. We do find, however, that this model is able to generate more symmetrical velocities than the SM89 model by relaxing the critical damping constraint. This introduces a different constraint: the model parameters must exist in a non-linear relation-

ship between b, k and duration in a way that avoids oscillation (Shaw and Chen 2019).

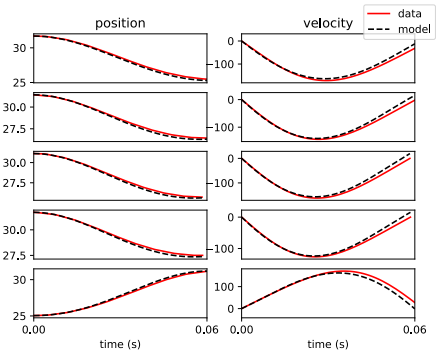


Figure 3: *Lip aperture trajectories and SINDy second-order model predictions. Model includes first-degree polynomials and physical constraints. The y-axis varies across each plot to fit the range of the data.*

If we relax the physical constraint $y \stackrel{!}{=} 1.00\dot{x}$ in $\ddot{x} = -by - kx$ then SINDy discovers the more complex model in (12):

$$\ddot{x} = -b(a - cx + d\dot{x}) - kx \tag{12}$$

Figure 4 shows example model predictions, with much improved fit between data and model. This comes at the cost, however, of adding significant complexity into the model.

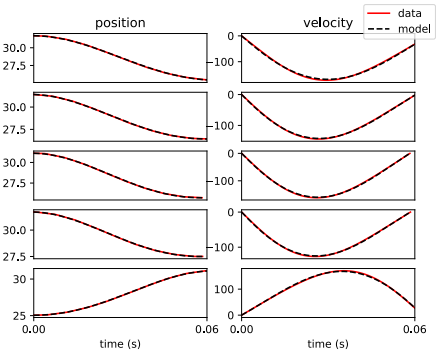


Figure 4: *Lip aperture trajectories and SINDy second-order model predictions. Model includes first-degree polynomials but no physical constraints. The y-axis varies across each plot to fit the data's range.*

5. Discussion and conclusion

This paper demonstrates how sparse symbolic regression can be used to identify dynamical principles of articulatory dynamics. The discovered models show a trade-off between simplicity and accuracy, from a simple first-order model that fits less accurately to a second-order model with no physical constraints that fits near-perfectly but is quite complex. In some cases, however, capturing the system's attractor dynamics may be more important than predicting trajectories, so the simpler models should not be immediately discounted. In future research, we will explore the discovered models via simulation to probe the dynamical principles they expose around the underlying system. In addition to this, we aim to test how well the discovered models generalise to different data sets. We note that the models should be treated with caution at this stage, as they are based on 29 trajectories of the same gesture from a single speaker, so these data may not be a good representation of all gesture types or speakers. This minimal proof-of-concept was driven by interpretability, but it clearly motivates extending this approach to a larger data set, which is the focus of ongoing research.

6. Acknowledgements

This research was supported by Arts and Humanities Research Council grant AH/Y002822/1.

7. 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.

Byrd, Dani and Elliot Saltzman (2003). "The elastic phrase: modeling the dynamics of boundary-adjacent lengthening". In: *Journal of Phonetics* 31.2, pp. 149–180.

Champion, Kathleen, Peng Zheng, Aleksandr Y. Aravkin, Steven L. Brunton, and J. Nathan Kutz (2020). "A unified sparse optimization framework to learn parsimonious physics-informed models from data". In: *IEEE Access* 8, pp. 169259–169271.

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.

Iskarous, Khalil, Jennifer Cole, and Jeremy Steffman (2024). "A minimal dynamical model of Intonation: Tone contrast, alignment, and scaling of American English pitch accents as emergent properties". In: *Journal of Phonetics* 101309.1–27.

Parrell, Benjamin, Antje Mefferd, Sarah Harper, Simon Roessig, and Doris Mücke (2023). "Using computational models to characterize the role of motor noise in speech: The case of amyotrophic lateral sclerosis". In: *Proceedings of the 20th International Congress of Phonetic Sciences* 988, pp. 878–882.

Pasquato, Mario, Mohamad Abbas, Alessandro A. Trani, Matteo Nori, Kwiecinski, Piero Trevisan, Vittorio F. Braga, Giuseppe Bono, and Andrea V. Macciò (2022). "Sparse identification of variable star dynamics". In: *The Astrophysical Journal* 930.161, pp. 1–13.

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

Schmidt, Michael and Hod Lipson (2009). "Distilling free-form natural laws from experimental data". In: *Science* 324, pp. 81–85.

Shaw, Jason A. and Wei-Rong Chen (2019). "Spatially conditioned speech timing: Evidence and implications". In: *Frontiers in Psychology* 10.2726, pp. 1–17.

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

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

Westbury, John R. (1994). *X-Ray Microbeam Speech Production Database User's Handbook*. Madison, WI: Waisman Center.

Spatiotemporal Features of Bilabial Geminate and Singleton Consonants in Italian

Francesco Burroni[1], Sireemas Maspong[1], Nicole Benker[1], Philip 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

Abstract

We investigate the production of Italian bilabial geminate and singleton consonants using electromagnetic articulography. The results reveal differences in closure duration, peak velocity, movement amplitude, and stiffness between geminate and singleton consonants. Timing analyses suggest earlier closure initiation and shorter trans-consonantal vocalic lags for geminates. Taken together, these findings suggest a distinct gestural specification for geminates, beyond longer gestural activations, and shed light on their acoustic manifestations, particularly the well-known shortening of vowels before geminate consonants and previously reported differences in how rate affects lags of geminates and singletons.

Keywords: articulation, geminates, Italian, speech production, electromagnetic articulography

1. Introduction

Speakers of Italian are known to employ consonantal duration contrastively, e.g., [pipa] "smoking pipe" vs. [pip:a] "pipsqueak". A substantial body of work has investigated the acoustic correlates of such singleton/geminate contrasts and found that the main difference between them lies in longer closure durations and shortening of vowels preceding geminates (cf. Di Benedetto et al., 2021 for a review), **Figure 1**.

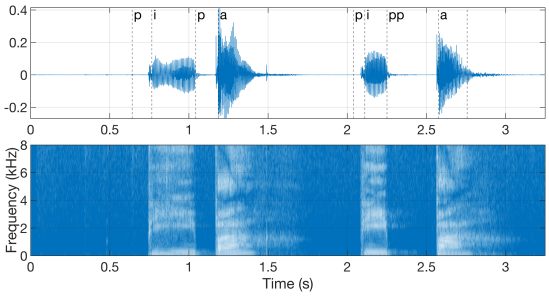


Figure 1: *Waveform and spectrogram of It. [pipa] vs. [pip:a].*

Less work has been dedicated to the kinematic properties underlying the acoustic features of geminate contrasts, and not all studies report compatible results. Most studies report robust differences in *closure duration* (Dunn, 1993; Gili-Fivela et al., 2007; Gili-Fivela & Zmarich, 2005; Zmarich et al., 2006, 2011). However, *peak velocity* has been reported to be sometimes higher for geminates (Gili-Fivela et al., 2007, 2015), equal for geminates and singletons (Gili-Fivela et al., 2007; Zmarich et al., 2006), and lower for geminates (Dunn, 1993; Smith, 1995). *Closure/release amplitude* has also been reported to be higher for geminates (Gili-Fivela et al., 2015; Zmarich et al., 2006), not different (Gili-Fivela & Zmarich, 2005), or even smaller for geminates (Dunn, 1993). Finally, *stiffness*, a parameter determining the time to target, has been reported to be consistently lower for geminates compared to singletons (Gili-Fivela & Zmarich, 2005; Zmarich et al., 2006). To our knowledge, no investigation of whether geminates exhibit more constricted targets has been conducted.

Similarly, the temporal dynamics inferred from kinematic studies are not always in agreement. Most studies report few, if any, consistent timing differences (Löfqvist, 2017; Zmarich et al., 2011). Among the most consistently reported timing differences is that geminate closures start earlier with respect to the preceding vowel (Celata et al., 2022; Dunn, 1993; Gili-Fivela et al., 2015; Smith, 1992), a pattern that offers a potential basis for acoustic shortening. However, even this timing difference has not been consistently observed (Gili-Fivela et al., 2007; Löfqvist, 2017). Other studies also reported shorter or equal V1-V2 intervals when the intervocalic is geminate vs. singleton (Smith, 1992), but other works have failed to replicate this finding or have reported longer V1-V2 intervals for geminates (Löfqvist, 2017; Zmarich et al., 2011). Finally, it has been suggested that differences between singletons and geminates may emerge or disappear under rate manipulations (Tilsen & Hermes, 2020; Zmarich et al., 2011); however, systematic studies of geminate kinematic and timing properties under rate manipulation based on a large pool of speakers are lacking. Given the open issues just outlined, we present the results of an electromagnetic articulography (EMA) study where we investigated the kinematic and timing properties of Italian (bilabial) geminates and singletons produced under rate manipulation by ten speakers of Italian.

1.1. Research Questions, Hypotheses, Predictions

The main research question we aim to answer is how Italian speakers produce the difference between geminate and singleton consonants in terms of the kinematic properties underlying their different acoustic outputs. Specifically, the question is investigated along the two lines discussed above:

1. How does the production of geminate and singleton consonants differ in terms of their *kinematic parameters* (closure duration, peak velocity, movement amplitude, stiffness, target)?
2. How does the productions of geminate consonants differ from that of singleton in terms of their *timing* to surrounding vocalic gestures? Do they affect the V1-V2 timing and, if so, how?

For RQ1, we can entertain two hypotheses, which have often been discussed in the literature regarding the nature of geminates (e.g., Di Benedetto et al., 2021; Dunn, 1993; Smith, 1992; Zeroual et al., 2015). The first hypothesis, H_{1A}, is that geminates are produced as longer versions of singletons with identical or similar kinematic parameters. The alternative hypothesis, H_{1B}, is that geminates are produced as different gestures compared to singletons with their own kinematic parameter specification. H_{1A} predicts that geminates, as longer versions of singletons, have longer durations and virtually identical kinematic parameters, except for those driven by longer durations; while H_{1B} predicts that geminates have longer durations as well as kinematic parameters, e.g., stiffness.

For RQ2, two influential timing regimes have been proposed in the literature (Smith, 1992). H_{1A} (V-V timing) is based on speech production models holding that vowels are timed to each other, while consonants are superimposed on them (Fowler, 1980; Öhman, 1966). Concretely, when geminates are superimposed on vowels the predictions are that the V_1-V_2 interval is stable and geminate closure can intrude earlier in the preceding vowel and later in the following. H_{1B} (V-C-V timing) is based on speech production models holding that consonants and vowels are directly timed to each other, e.g., Articulatory Phonology (Browman & Goldstein, 1989). Concretely, when geminates are timed to the preceding (and following) vowel(s), depending on specific coordination patterns, they can affect the duration of the V_1-V_2 interval, Figure 2.

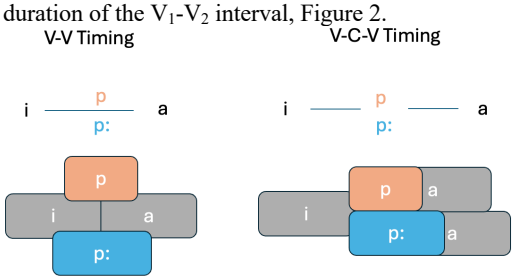


Figure 2: *Illustration of geminate V-V timing and V-C-V timing.*

2. Methods

2.1. Participants, materials, and procedures

We collected simultaneous audio and EMA (3D Carstens AG501) data from 10 native Italian speakers, speaking Central and Southern varieties (south of the Rimini-La Spezia line) where geminates are known to be clearly distinguished (Mairano & De Iacovo, 2020). Three reference sensors were positioned on the left and right mastoid process and on the nasion to correct for head movement. Additionally, three tongue sensors were placed: one on the tongue tip (TT), ~5 mm posterior to the tongue apex; another on the tongue dorsum (TD), positioned as close to the terminal sulcus as comfortable for participants; and the last on the tongue body (TB), at the midpoint between the TT and TD sensors. Two sensors below the inferior left incisor and above the superior left incisor were used to track mandibular and maxillary movement. Finally, two sensors were placed on the upper (UL) and lower lip (LL) vermilion borders to track lip movements. In this paper, we focus on the TT,TB,TD and UL,LL sensors. Participants were instructed to produced six disyllabic ˈVC(ː)V pseudowords containing all singleton and geminate Italian bilabial consonants: [ipa, ipːa, iba, ibːa, ima, imːa]. We refer to /i/ as V1 and /a/ as V2. A high to low vowel transition was chosen to maximize tongue vertical movement and facilitate landmarking. Bilabial consonants were chosen to avoid competing demands on tongue movement from consonants and vowels and obtain precise timing estimates. Target words were embedded in the sentence [dika __ due vɔlte] 'Please say __ twice'. In trials participants were cued to produce at 5 rates 'very slow', 'slow', 'normal', 'fast', 'very fast', to introduce variability in rate . Each word was repeated 12 times at each rate. We collected 360 tokens per speaker. Following the exclusion of trials that contained disfluency or equipment malfunctions, we retained 3,593 tokens for analyses.

2.2. Data processing and statistical analyses

Bilabial consonants' closure (CLO), plateau, and release (REL) phases were identified using a lip aperture (LA) time series. LA was defined as the 3D Euclidean distance between the LL and UL sensors. Vocalic gestures were identified on the basis of the first principal component (PC) of tongue movement obtained by entering three-dimension movement components of the TT, TB, and TD sensors in a PC analysis. The 1st PC accounts for 92% of variance on average. Consonantal and vocalic gesture landmarks were identified using a 20% threshold on peak velocity in the vicinity of acoustic landmarks obtained from forced alignment. From landmarking we extracted the following 7 variables (Figure 3): (1) Duration of the closure phase (CLO Dur + Plateau Dur); (2) CLO amplitude; (3) absolute peak velocity of CLO; (4) "Stiffness" of CLO (ratio of absolute peak velocity and movement amplitude); (5) LA minimum, as a proxy for constriction target; (6) V_1-CLO lag; (7) V_1-V_2 lag.

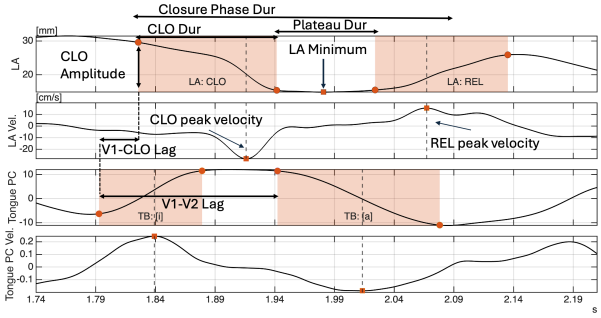


Figure 3: *Example of velocity-based landmarking of LA and Tongue PC with derived measurements.*

All dependent variables were analyzed with linear mixed effect regression models using the *fitlme*() function in MATLAB. Fixed effects were rate in z-scored phonemes per second (continuous variable, μ=10 phon/s, σ=3.8 phon/s), geminate status (categorical, with reference as geminate), and their interaction. Maximal random effect structures (with intercepts and slopes) for subject session and voicing/manner, *i.e.*, whether the consonant is [p], [b], or [m], and repetition number were also included. Model selection was accomplished using log-likelihood ratio tests performed with the *compare*() function in MATLAB. We compared models stepwise by first eliminating the interaction term, then the geminate term.

3. Results

3.1. Duration and kinematic parameters

For closure phase duration, we found that geminate tokens have a longer closure phase 193 ms (SE 13.6), while singleton tokens have a shorter closure phase –56 ms, (SE 9.1). Closure phases also shortens as rate increase by –58 ms (SE 9.2) per one z-score unit increase in rate for geminates, but less for singletons where the effect is estimated at –19 ms, as there is an interaction term rate×singleton estimated at +39 ms (SE 6). Model outputs overlaid over individual observations are presented in Figure 4.
For closure amplitude, we found that geminates have a greater movement amplitude estimated at 11.2 mm (SE 0.7), singletons movement amplitude is –1.05 mm (SE 0.14) less wide, and as rate increases amplitude of movement also increases by 1.16 mm (SE 0.19) per one z-score increase in rate, Figure 5.
For peak velocity, we found that geminate peak velocity is faster, estimated at 18.8 cm/s (SE 1.6), singleton peak velocity is slower by –0.75 cm/s (SE 0.34). Peak velocity also increases by 3.3 cm/s (SE 0.45) per one z-score increase in rate, Figure 6.
For stiffness, we found that geminates have a lower stiffness value estimated at 17 s^{-1} (SE 1.12), singleton stiffness is higher

by 1.04 s^{-1} (SE 0.32). Stiffness also increases by 0.9 s^{-1} (SE 0.4) per one z-score increase in rate, **Figure 7**.

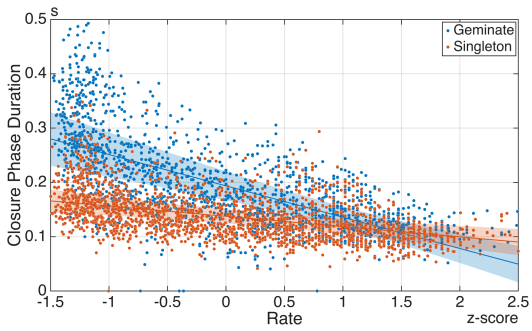


Figure 4: *Geminate/singleton closure phase duration.*

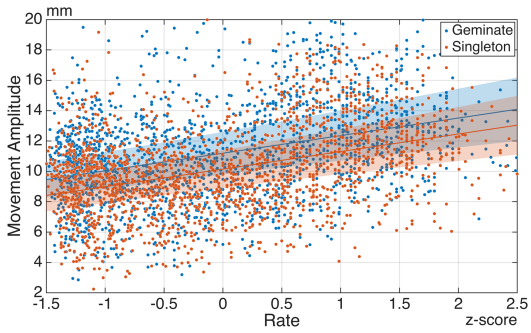


Figure 5: *Geminate/singleton movement amplitude.*

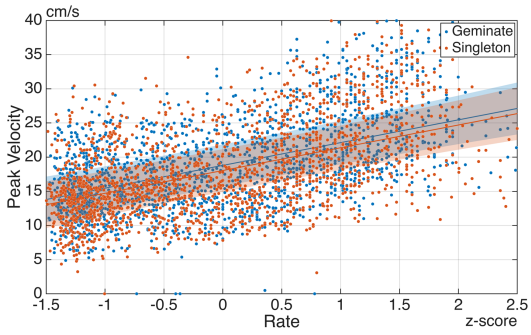


Figure 6: *Geminate/singleton peak velocity.*

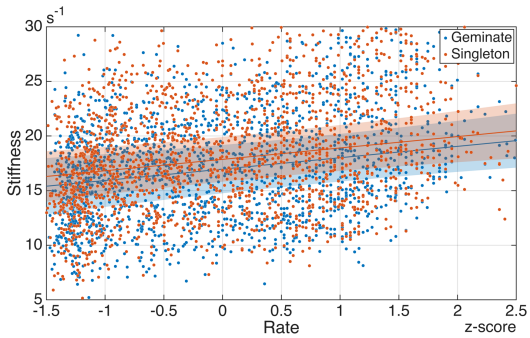


Figure 7: *Geminate/singleton stiffness.*

For minimum LA value, a proxy for target, we found that geminates have a more constricted minimum LA value estimated at 16.8 mm (SE 0.8), singletons minimum LA is higher, *i.e.*, less constricted, by 0.8 mm (SE 0.09). LA minimum values also increases by 0.4 mm (SE 0.05) per one z-score increase in rate, **Figure 8**.

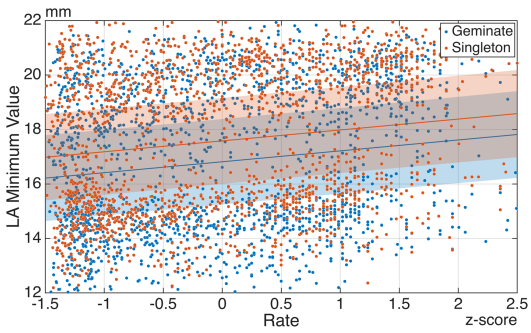


Figure 8: *Geminate/singleton LA minimum.*

3.2. Timing

For the V1-CLO onsets lag, we found that geminates are associated with a shorter V1-CLO lag estimated at 123 ms (SE 11), singletons display a longer lag with an effect estimated at 43 ms (SE 7.4). The V1-CLO lag decreases by –109 ms (SE 11.3) and it does so faster for singletons with an additional –33 ms (SE 5.3) per one z-score increase in rate, **Figure 9**.
For the V1-V2 onsets lag, we also found that geminates are associated with a shorter V1-V2 lag estimated at 203 ms (SE 9), the singletons V1-V2 lag is +23 ms (SE 3) longer. The V1-V2 lag decreases by –103 ms (SE 9.2) and it does so faster for singletons with an additional –27 ms (SE 3.9) per one unit increase in rate, **Figure 10**.

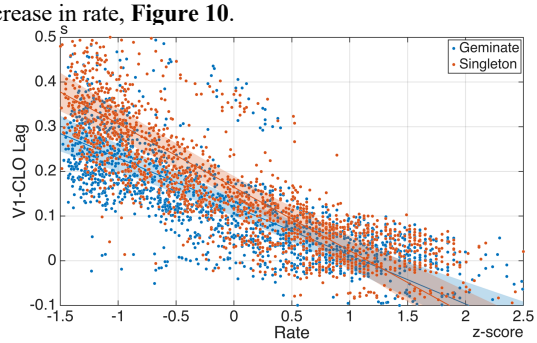


Figure 9: *Geminate/Singleton V1-CLO lag.*

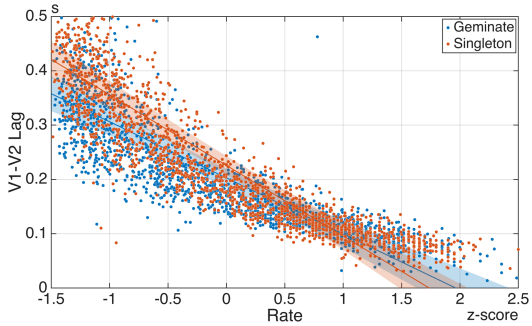


Figure 10: *Geminate/Singleton V1-V2 lag.*

4. Discussion and conclusion

Our results suggest that Italian bilabial geminates differ from singletons along a variety of kinematic, durational, and intra-/intergestural timing properties. With respect to duration, our results confirm previous work in showing that geminates are produced with longer closure phases. Singletons, even at very slow rates, do not reach comparably high durational values and generally have more "constrained" closure phase duration across rates (Tilsen & Hermes, 2020).

With respect to kinematic parameters, we found that, even when produced at various speech rates, geminates are produced with wider articulator movements, slightly faster peak velocity, slightly lower stiffness, and more constricted targets. Given the differences in kinematic parameters, it seems possible that geminates are not singletons with longer activation intervals (Gafos & Goldstein, 2012), but a distinct gestural category characterized by more extreme targets (Löfqvist, 2005), and, arguably, slightly lower "stiffness" and slightly higher peak velocities. The idea that geminates are not produced by speakers simply as longer versions of singletons is also supported by their different timing regimes to surrounding vowels.

With respect to timing patterns, our main findings are that the lag between V1 and geminates CLO is shorter for geminates, suggesting that geminate closure is initiated earlier with respect to V1 (Celata et al., 2022; Dunn, 1993; Smith, 1995). At first glance, this timing regime may suggest V-V timing for Italian, in which consonants are overlaid onto vocalic intervals. However, we also observed that V1-V2 intervals are not identical across geminate and singleton consonants. The V1-V2 interval is shorter when the intervening consonant is geminate compared to singleton. This fact suggests that V1-V2 lags are not constant, contra the predictions of V-V timing. Our data suggests not only that geminate closure gestures start earlier, but that a vowel following a geminate also starts earlier compared to when it starts after a singleton. That is, in the presence of a geminate, the following vowel is anticipated like the closure with respect to the preceding vowel, possibly because the two are timed to each other. Earlier initiations of the closure and second vowel provide a potential basis for the acoustic shortening effect observed for pre-geminate vowels in Italian. Vowels preceding geminates are shorter because they are likely truncated by the earlier initiation of a following articulatory gesture (Cho, 2006). Additionally, as a mirror image for intragestural closure phase durations, the V1-CLO and V1-V2 lags are more constrained across rates suggesting that they may have ceiling values at slower rates for geminates, as hypothesized for the singleton CLO-REL lag (Tilsen & Hermes, 2020). Thus, when Italian speakers produce bilabial geminates at slower rates, they can more freely extend the closure phase duration, but they cannot in the same way extend the lags with surrounding vowels. Conversely, for singletons, speakers cannot extend the duration of the closure phase, but they can extend the lags to surrounding vowels. Taken together, our findings suggest that the acoustic manifestations of Italian bilabial geminate consonants are rooted in a distinct spatiotemporal articulatory profile. Their gestural specification embraces spatial characteristics, deriving from a slightly different set of kinematic parameters, and temporal characteristics, deriving from different timing regimes to surrounding vowels, as well as how these are affected by rate. Future work should explore the kinematic and timing parameters of lingual geminates which share the same articulator with vowels and in languages that lack pre-geminate vowel shortening, e.g., Japanese.

5. References

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251.

Celata, C., Meluzzi, C., & Bertini, C. (2022). Acoustic and kinematic correlates of heterosyllabicity in different phonological contexts. *Language and Speech*, *65*(3), 755–780.

Cho, T. (2006). Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English. *Laboratory Phonology*, *8*, 519–548.

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.

Dunn, M. H. (1993). *The phonetics and phonology of geminate consonants: A production study*. Yale University.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*(1), 113–133.

Gafos, A., & Goldstein, L. (2012). Articulatory representation and organization. *The Oxford Handbook of Laboratory Phonology*, 220–231.

Gili-Fivela, B., Iraci, M. M., Grimaldi, M., & Zmarich, C. (2015). Consonanti scempie e geminate nel Morbo di Parkinson: La produzione di bilabiali. *Studi AISV*, 289–312.

Gili-Fivela, B., & Zmarich, C. (2005). Italian geminates under speech rate and focalization changes: Kinematic, acoustic, and perception data. *Interspeech 2005*, 2897–2900.

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*, 469–472.

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. (2017). Articulatory coordination in long and short consonants: An effect of rhythm class? In *The Phonetics and Phonology of Geminate Consonants* (pp. 118–129).

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.

Öhman, S. E. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*(1), 151–168.

Smith, C. L. (1992). *The timing of vowel and consonant gestures*. Yale University.

Smith, C. L. (1995). Prosodic patterns in the coordination of vowel and consonant gestures. *Papers in Laboratory Phonology IV, Phonology and Phonetic Evidence. CUP*, 205–222.

Tilsen, S., & Hermes, A. (2020). Nonlinear effects of speech rate on articulatory timing in singletons and geminates. *12th International Seminar on Speech Production*.

Zeroual, C., Hoole, P., Gafos, A. I., & Esling, J. H. (2015). Gestural coordination differences between intervocalic simple and geminate plosives in Moroccan Arabic: An EMA investigation. *International Congress of Phonetic Sciences*. https://api.semanticscholar.org/CorpusID:44392979

Zmarich, C., Gili-Fivela, B., Perrier, P., Savariaux, C., & Tisato, G. (2006). Consonanti scempie e geminate in italiano: Studio acustico e cinematico dell'articolazione linguale e bilabiale. *Atti Del III Convegno Nazionale Dell'Associazione Italiana Di Scienze Della Voce (AISV)*.

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*, 401–404.

A dynamic neural field model of vowel diphthongisation

Sam Kirkham[1], Patrycja Strycharczuk[2]

[1]*Lancaster University, UK*
[2]*University of Manchester, UK*
s.kirkham@lancaster.ac.uk, patrycja.strycharczuk@manchester.ac.uk

Abstract

We advance a computational model of vowel diphthongisation that situates phonological representations in dynamic neural fields (DNFs), which represent the time-varying activation of neural populations that are sensitive to a given phonetic parameter range. We model all long vowels as two separate inputs to the DNF, with input timing governed by a coupled oscillator model that generates an anti-phase relationship between inputs. The location of time-varying maximum activation in the DNF forms a noisy dynamic target, which is used as input to a task dynamic model of gestural coordination. We find that spatial characteristics of long vowels are well captured by the model, which exhibits gradient variation between monophthongs and diphthongs. We also show that a simplified model of production/perception can simulate changes in a speaker's phonological planning representations, which could represent a mechanism behind sound change if transmitted across a community.

Keywords: Articulatory Phonology, Task Dynamics, Dynamic Field Theory, computational modelling, vowels

1. 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, as well as diphthongisation of tense monophthongs, such as FLEECE and GOOSE (Hughes, Trudgill, and Watt 2012). Some speakers even alternate between such variants, such as producing variably diphthongal or monophthongal vowels. The issue of variable diphthongisation also underpins accounts of diachronic change, such as the diphthongisation of Middle English /i/ and /u/ into present-day /aɪ/ and /aʊ/, as a consequence of the English Great Vowel Shift (Jespersen 1909).

In Strycharczuk et al. (submitted), we account for the gradient nature of diphthongisation by proposing a compositional two-target model for all long vowels (following precedents in Labov, Ash, and Boberg 2006; Popescu and Chitoran 2022). In this view, a short monophthong is short because it has a single target, while 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 as a long monophthong (two targets), but has different parameters for each of the targets, thus yielding movement from the first target to the second. However, this model does not contain appropriate mechanisms that would help to explain observed variability in vowels, such as the role of perceptually-driven change and the mechanisms behind variability in an individual speaker. One possibility is that each speaker has a single target for the component gestures, but that over time a community drifts towards a new set of targets. This hypothesis is untenable, as we know that speakers can also be highly variable. An alternative is that an individual speaker has a distribution of targets, which would facilitate an account of observed variability. But where do these distributions originate and how do they undergo change?

We outline a solution by grounding phonological representations in a dynamical planning field. Specifically, we use the mathematical and conceptual insights of dynamic field theory (DFT) (Schöner, Spencer, and The DFT Research Group 2016), which have proven to be a versatile tool for dynamical models of phonological planning (Kirov and Gafos 2007; Tilsen 2007; Roon and Gafos 2016; Tilsen 2019; Harper 2021; Shaw and Tang 2023; Stern and Shaw 2023). A dynamic neural field (DNF) model situates phonological planning in an activation field over a phonetic parameter range. A dynamical equation specifies the evolution of field activation until some value reaches a threshold, which is selected as the parameter value for speech production. We then model production and perception as inputs to the field, allowing us track how the field develops over real-time speech planning, as well as over longer timescales. The following model is inspired by integrative dynamical models of timing, planning and execution (Tilsen 2018; Tilsen 2019), as well as by the proof-of-concept DFT model of sound change in Kirov and Gafos (2007).

2. Model architecture

2.1. Dynamic neural field model

A phonological planning representation is modelled as a dynamic neural field, which evolves according to (1) (Schöner, Spencer, and The DFT Research Group 2016). τ dictates the rate of field evolution, $-u(x,t)$ is time-dependent activation at each field site x, h is the resting level of the neural field, $s(x,t)$ represents an input to the field, and $\xi(x,t)$ is Gaussian noise scaled by a factor q.

$$\tau \dot{u}(x,t) = -u(x,t) + h + s(x,t) + \int k(x-x')g(u(x',t))dx' + q\xi(x,t) \quad (1)$$

An input $s(x,t)$ represents any task-specific input, such as phonological planning units or perceptual input, and is modelled in (2) as a Gaussian distribution over a parameter x with amplitude a, centroid p and width w. A model can have multiple inputs, which are summed as $s_1(x,t) + s_2(x,t) + s_n(x,t)$.

$$s(x,t) = \sum_i a_i \exp\left[-\frac{(x-p_i)^2}{2w_i^2}\right] \quad (2)$$

The interaction kernel $k(x - x')$ in (3) defines excitatory and inhibitory forces across the DNF. Each field location only contributes to above-threshold activation when it exceeds a threshold of $u = 0$. Interaction is excitatory for nearby locations and inhibitory for distal locations. c_{exc}, σ_{exc} are the mean and standard deviation of the excitatory component, while c_{inh}, σ_{inh} are the mean and standard deviation of the inhibitory component. c_{glob} is a global inhibition constant.

$$k(x - x') = \frac{c_{exc}}{\sqrt{2}\pi\sigma_{exc}} \exp\left[-\frac{(x - x')^2}{2\sigma_{exc}^2}\right] - \frac{c_{inh}}{\sqrt{2}\pi\sigma_{inh}} \exp\left[-\frac{(x - x')^2}{2\sigma_{inh}^2}\right] - c_{glob} \quad (3)$$

The interaction kernel is gated by a sigmoidal function $g(u)$, where β is the slope of the sigmoid and α is a threshold, typically set to $\alpha = 0$, whereby only activation values above zero contribute to supra-threshold activation.

$$g(u) = \frac{1}{1 + \exp(-\beta(u - \alpha))} \quad (4)$$

2.2. Coupled oscillator model of gestural timing

We model phonological planning as separate planning inputs $s_{nuc}(x, t)$, $s_{glide}(x, t)$ for the nucleus and offglide. The relative timing of these inputs is determined via the coupled oscillator model in (5) (Tilsen 2018). Φ_{ij} is the relative phase between oscillators i, j, such that $\Phi_{ij} = \theta_i - \theta_j$. C_{ij} is a matrix of coupling strengths between oscillators i, j, where $C_{ij} > 0$ is in-phase and $C_{ij} < 0$ is anti-phase.

$$\dot{\theta}_i = 2\pi f_i + \Sigma_j C_{ij} \sin(\Phi_{ij}) \quad (5)$$

We model all planning units with the same oscillator frequency f = 4 Hz and each unit lasts for 200 ms. If two vowel planning units of 200 ms are coupled anti-phase then the offglide will begin 100 ms after the nucleus. This does not mean, however, that the period of activation will be 300 ms, as there is a time lag between an input to the DNF and activation reaching the threshold. Above-threshold activation can also persist after an input is removed, due to stability-promoting mechanisms in the model. We ensure realistic vowel durations by setting input amplitudes such that activation relaxes to resting level shortly after an input is removed. While we believe that the timing of gestural onsets via coupled oscillators is neurally plausible, the notion of fixed input durations is likely not, so this represents a simplifying heuristic in lieu of a more realistic mechanism, such as feedback-induced gestural suppression (Tilsen 2019).

2.3. Task dynamic model

The DNF governs gestural selection, activation durations, and time-varying gestural targets. We model gestural dynamics using the model in (6) from Saltzman and Munhall (1989), where m is mass, b is a damping coefficient, k is a stiffness coefficient. The task dynamic literature conventionally defines $m = 1$ and $b = 2\sqrt{m}k$, which makes (6) a critically damped oscillator (see Iskarous 2017 for an accessible overview of this model).

$$m\ddot{x} + b\dot{x} + k(x - T(t)) = 0 \quad (6)$$

Gestures are commonly represented by a single target T, but the DNF produces time-varying activations across a parameter range, which represent a dynamic target $T(t)$. Tilsen (2019) proposes a DNF model with dynamic targets, whereby an activation-weighted target supplants the gestural blending mechanism of Saltzman and Munhall (1989). In our study, the target simply tracks the location of peak activation. This enforces stricter selection dynamics, as sudden changes in the location of peak activation results in sudden changes in the target.

The presence of neural noise in the DNF means that the location of peak activation is often a noisy function of time, so how do we avoid overly noisy gestural trajectories? The key concept is that the time-varying location of peak activation is a dynamic input $T(t)$ to the model in (6), not the actual articulatory movement trajectory. The stiffness term k acts as a restoring force that governs the acceleration of the system. Lower values of k constrain movement between dynamic target values, essentially acting as a low-pass filter that forces smoothness on trajectories. Importantly, this is not a form of ad-hoc smoothing, but inherent to the dynamics of the system, allowing smooth gestural trajectories to emerge from noisy neural outputs.

2.4. Computational implementation

All computational models in this paper were implemented in Python 3.9.13, with numerical integration computed using `scipy.integrate.solve_ivp`. Numerical parameters are as follows: DNF [$x = [-10, 10]$, $\tau = 50$, $h = -2$, $\xi = \mathcal{N}(0, 1)$, $q = 3$, $\Delta t = 0.001$, $\Delta x = 0.1$]; kernel: [$c_{exc} = 1, c_{inh} = 0.5, c_{glob} = 0.1, \sigma_{exc} = 1, \sigma_{inh} = 3$], sigmoid: [$\alpha = 0, \beta = 1.5$]; coupled oscillator: [$\Delta t = 0.001, f = 4$]; task dynamics: [$\Delta t = 0.001$, $m = 1$, $k = 2/\Delta t$]. Parameter values encode relative relationships between elements and the specific values are not integral to the model.

3. Simulation results

3.1. Long monophthongs vs. diphthongs

Figure 1 shows example DNFs for three cases: (1) a long monophthong with two identical targets, p = [0,0]; (2) a diphthong with two different targets, p = [3,0]; (3) a diphthong with a bigger distance between two targets, p = [5,0]. The first target has amplitude $a = 3$ and the second $a = 6$ to represent the difference in relative blending weight in favour of the offglide in traditional task dynamic models. The parameter range represents an abstraction for Tongue Body Constriction Location (TBCL), where 0 is a palatal vowel and 5 is a pharyngeal vowel (the parameter range is purely heuristic for the purposes of illustration). In the top row, the left panel shows a single peak: the second input is at the same location as the first, thereby boosting the peak to a higher activation level. The middle panel shows the emerge of two peaks, which briefly overlap, causing a sudden change in the location of maximum activation. The right panel shows a similar dynamic, but the first input sits at a higher value on the parameter range, resulting in a larger difference in the location of peak activation between onset and offset.

The second row of Figure 1 shows the location of peak activation on the parameter axis (the noisy field means there are minor jumps in this value between time-steps). The third and fourth rows show the output of task dynamic simulations, with the dynamic target as time-varying input (initial position $x = -1$; initial velocity = 0 for all examples). Note that the position and velocity trajectories are moderately smooth, with some minor perturbations in the velocity signal. This demonstrates that the distinction between a long monophthong (single velocity minimum) and a diphthong (two velocity minima) can be captured by the model.

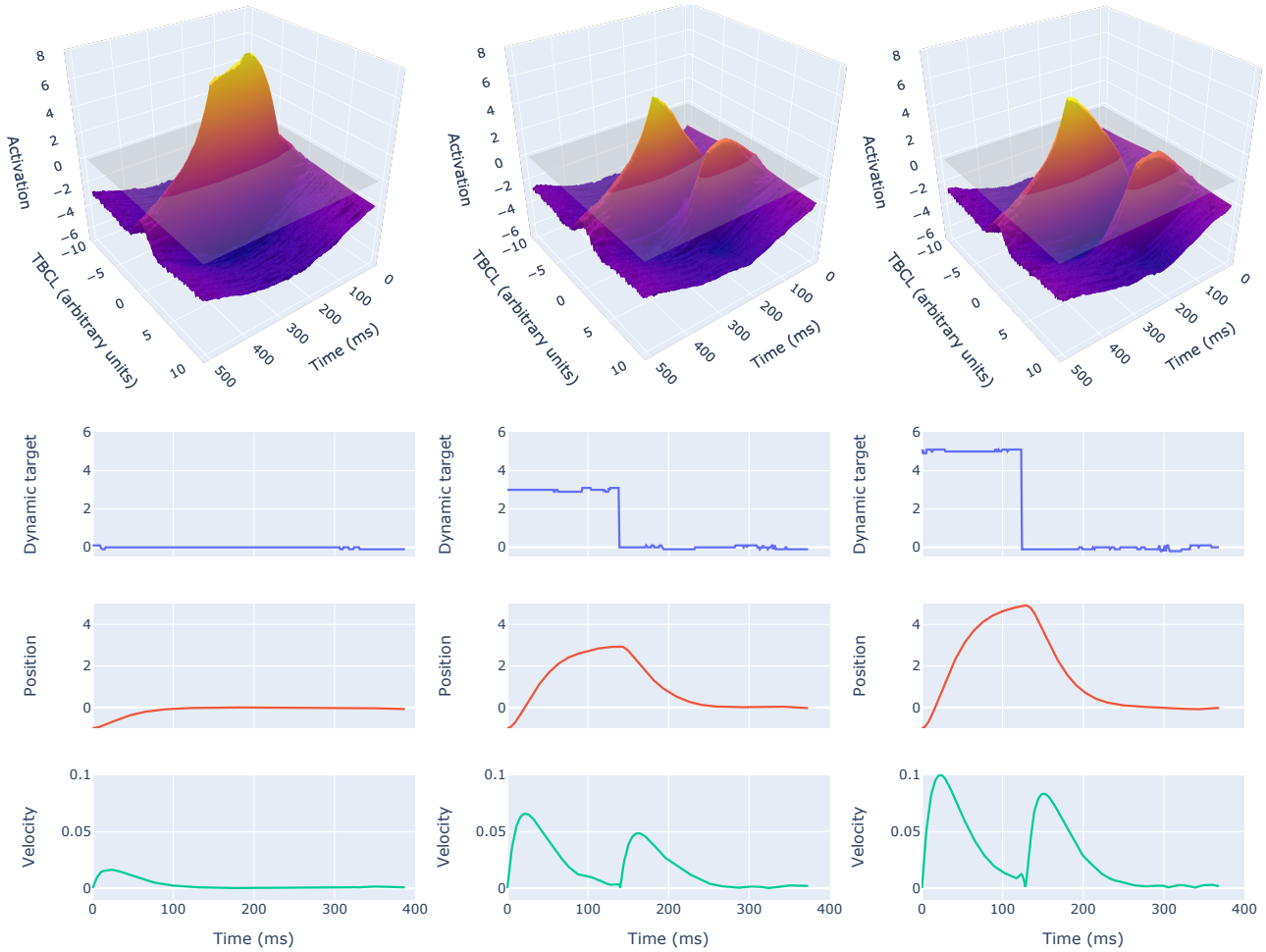


Figure 1: *ROW 1: DNFs for three vowels with increasing degrees of diphthongisation from left-to-right (grey plane = threshold). ROW 2: time-varying location of peak activation in each DNF. ROWS 3 & 4: Task dynamic simulation based on dynamic targets from each DNF. Time in rows 2–4 corresponds to the interval between the onset and offset of supra-threshold activation in each field.*

3.2. Production-perception model

We now present a model of how a speaker's phonological planning representation could undergo change from a long monophthong to a diphthong. This is a highly simplified model of production-perception inspired by Kirov and Gafos (2007) in which two speakers (A and B) interact. Specifically, speaker A produces a long monophthong with two identical targets. They then perceive speaker B producing the same vowel, but with a slightly different phonetic target for the nucleus. This represents perceptual input to speaker A's DNF, which changes their memory trace for the next production to a minor degree. This process repeats, with speaker A producing a vowel, perceiving speaker B's vowel, and so on. This is obviously a highly idealised model of interaction, as the influence is unidirectional (speaker B influences speaker A, but speaker A does not influence speaker B) and the only variation in speaker B's production is due to the addition of random noise added to their target value.

A long vowel is comprised of two inputs: $s_{nuc}(x,t)$ and $s_{glide}(x,t)$. We keep $s_{glide}(x,t)$ constant across production-perception loops, but vary $s_{nuc}(x,t)$ according to (7), where α and γ are weights for the respective task and perceptual inputs. Nucleus and glide both begin with $p = 0, w = 0.7$, with input amplitudes of $a = 3$ (nucleus) and $a = 6$ (offglide). Across production-perception loops, the current $s^i_{nuc}(x,t)$ is:

$$s^i_{nuc}(x,t) = \alpha s^{i-1}_{nuc}(x,t) + \gamma s_{perception}(x,t) \quad (7)$$

$s_{perception}(x,t)$ is defined as in equation (2) for $s(x,t)$, with $a = 0.3$, $w = 0.7$, except p is calculated as:

$$p = \arg\max_x u(x,t) + bias + q\xi \quad (8)$$

where $\arg\max_x u(x,t)$ is the TBCL parameter corresponding to the location of maximum activation (sampled at t = 100), $bias$ is a numerical value representing the difference between speaker A's target and the perceived phonetic target from speaker B (here $bias$ = 1.5), and q is a weighting factor that scales Gaussian noise ξ in the range [0,1]. The task input $s_{nuc}(x,t)$ is weighted by $\alpha = 0.99$, representing very slow memory decay, and the $s_{perception}(x,t)$ input is weighted by $\gamma = 0.2$. Higher values of γ increase the influence of the perceptual input, resulting in faster change over repeated loops.

The production-perception loop was run for 150 iterations and the resulting activation distributions at several iteration

steps are shown in Figure 2. After a number of interactions with this 'biased' speaker B, speaker A's activation field for the nucleus shifts away from the initial state towards a new peak. Notably, the offglide peak does not change very much at all, showing that this target remains stable. The nucleus, however, undergoes change, with the resulting vowel being gradually more diphthongal because the centroid of the nucleus distribution increasingly diverges from the offglide as the iterations increase.

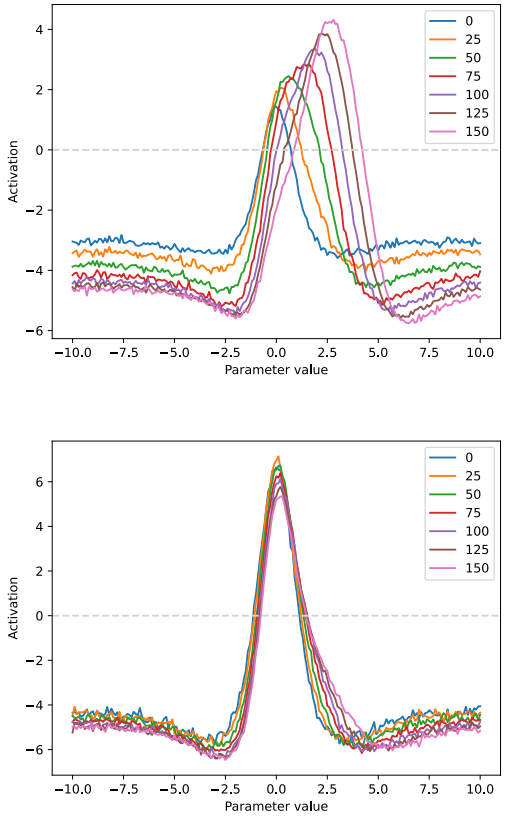


Figure 2: *Activation distributions at selected steps of the production-perception loops for nucleus target sampled at* $t = 100$ *(top) and offglide target at* $t = 300$ *(bottom).*

4. Discussion and conclusion

In summary, we model gestural selection, activation and articulatory dynamics using a combination of dynamic field theory, coupled oscillators and task dynamic models. This allows us to pose specific mechanistic connections between different components of the model, which yields behaviourally-realistic articulatory trajectories for long monophthongs and diphthongs, grounded in neurally-plausible dynamical mechanisms. We also use the same mathematical and conceptual language to propose a mechanism for variation and change in the phonological representations of individual speakers, thereby identifying a clear link between short-term synchronic variation and medium-term change in the diphthongisation of long vowels.

In terms of future research, the model assumes that gestural parameters, such as TBCL, are directly retrievable in perception. While speakers can undoubtably infer articulatory gestures from acoustics, the mapping is unlikely to be linear or perfect and a more realistic model requires a perceptual-acoustic field that projects to a tract variable field. Second, our model of between-speaker interactions is highly idealised and our future research aims to develop more complex models of interaction between small groups of speakers. Finally, while our DNF claims to be a neural model, we make no claims about cortical or subcortical localisation. Instead, the DNF is an abstraction that models the functional behaviour of a neural population, which may actually be distributed over different areas of the brain (Schöner, Spencer, and The DFT Research Group 2016).

5. Acknowledgements

This research was supported by Arts and Humanities Research Council grants AH/S011900/1 and AH/Y002822/1.

6. References

Harper, Sarah Kolin (2021). "Individual differences in phonetic variability and phonological representation". PhD thesis. Los Angeles, CA: University of Southern California.

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.

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.

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.

Labov, William, Sharon Ash, and Charles Boberg (2006). *The Atlas of North American English*. Berlin: Mouton de Gruyter.

Popescu, Anisia and Ioana Chitoran (2022). "Linking gestural representations to syllable count judgements: A cross language test". In: *Laboratory Phonology* 13.1, pp. 1–48.

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.

Stern, Michael C. and Jason A. Shaw (2023). "Neural inhibition during speech planning contributes to contrastive hyperarticulation". In: *Journal of Memory and Language* 132.104443, pp. 1–16.

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

Tilsen, Sam (2007). "Vowel-to-vowel coarticulation and dissimilation in phonemic-response priming". In: *UC Berkeley Phonology Lab Annual Report* 3.1, pp. 416–458.

— (2018). "Three mechanisms for modeling articulation: selection, coordination, and intention". In: *Cornell Working Papers in Phonetics and Phonology*, pp. 1–49.

— (2019). "Motoric mechanisms for the emergence of non-local phonological patterns". In: *Frontiers in Psychology* 10, pp. 1–25.

Sonority patterns and onset cluster production in Mandarin

Xuejing Chen[1], Rachid Ridouane[1], Pierre Hallé[1]

[1]*Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), 4 Rue des Irlandais, 75005 Paris*

xuejing.chen@sorbonne-nouvelle.fr, rachid.ridouane@sorbonne-nouvelle.fr, pierre.halle@sorbonne-nouvelle.fr

Abstract

*This study examined the role of the Sonority Sequencing Principle (SSP) in the production of word-initial clusters by Chinese speakers. We conducted an imitation experiment in which Chinese participants had to imitate "model" speech stimuli of the form C1C2a or C1əC2a, with 3 types of sonority profile for C1C2: rising (e.g., kla), plateau (e.g., kpa), falling (e.g., lka). If the SSP influences the production of these clusters, one would expect a higher incidence of vowel insertion for more marked sonority profiles. Our results are consistent with this prediction: more epenthetic vowels were produced within more marked C1C2 clusters, suggesting SSP effects in their production. The acoustic characteristics of the epenthetic vowels suggest they were all the more "intended" (*i.e.*, targeted) that the model clusters were marked. This pattern suggests that the observed SSP effects in terms of incidence of vowel insertion do not solely reflect perceptual effects during the imitation task.*

Keywords: Sonority Sequencing Principle, Onset clusters, Chinese speakers' productions, Epenthetic schwas

1. Introduction

The Sonority Sequencing Principle (SSP) might explain a putatively universal preference for well-formed over ill-formed syllables with respect to onset sonority profile (Clements 1990): a well-formed or more acceptable syllable is one in which the sonority profile increases maximally from the onset to the nucleus, then decreases minimally from the nucleus to the coda. For #C1C2V sequences, the SSP posits that the onset cluster #C1C2 is well-formed if its sonority profile rises monotonically from the beginning to the nucleus, and the worst-formed if its sonority profile is falling. Thus, a rising sonority profile is better formed than a plateauing profile, which is in turn better formed than a falling profile. The sonority profile of onset clusters can be inferred by reference to Clements' (1990) scale of sonority, which is the most commonly used: vowel > glide > liquid > nasal > obstruent. Hence, a structure like *bla* (rising sonority profile) is well-formed and universally preferred over e.g. *lba* (falling sonority profile), which is ill-formed. Even though many languages allow onset clusters that violate the SSP, such as Russian permitting falling profiles like /lp/, or Hebrew allowing plateau profiles like /kp/ (see Yin et al. 2023), universal regularities governed by the SSP appear in the distribution of onset clusters (as already found by Greenberg 1978).

Now, what are the perceptual consequences of the SSP?

Several studies have demonstrated that phonotactically illegal clusters are not perceived faithfully (Davidson & Shaw 2012; Dupoux et al. 1999; Hallé et al. 1998): they tend to be perceptually "repaired." A common perceptual repair for illegal clusters is epenthetic-vowel insertion: C1C2 > C1vC2 (e.g., Dupoux et al. 1999: *ebzo* > *ebuzo*). In Dupoux *et al.*'s formulation, subjects perceive an illusory vowel /u/ in *ebzo*. In addition to cluster grammaticality, sonority-related restrictions dictated by the SSP also trigger epenthetic-vowel perception. This was shown in a series of perception studies by Berent and colleagues (e.g., Berent et al. 2007, 2008, 2012): for highly marked onsets in terms of the SSP, such as in *lbif*, listeners tended to perceive monosyllabic *lbif* as dissyllabic, suggesting they perceive *lbif* as *lə.bif*. The presumed misperception of an epenthetic vowel was found to be more likely for more marked clusters in terms of SSP (*lbif* > *bdif* > *bnif*). As a consequence, discrimination between CCif and CəCif was more difficult for more marked CC clusters, independently of the fact that the CCs used (e.g., *bn*, *bd*, *lb*) all are banned in English as onsets. Therefore, the SSP seems to be at work to determine epenthetic-vowel perception in addition to phonotactic violation. Similar findings hold for adults from various L1s, including some that ban clusters altogether (French, Hebrew, Spanish, but also Korean, or Chinese: Berent et al. 2008, 2012, 2013, 2016; Maïonchi-Pino et al. 2015) as well as for French children aged 8-12 years (Maïonchi-Pino et al. 2015). Gomez *et al.*'s (2014) fNIRS study suggests that the preference for *bl* over *lb* or *bd* already appears at birth. Moreover, sensitivity to SSP-defined well-formedness may not be specific to human listeners according to Santolin *et al.* (2023), who found it in rats and argue that "sensitivity to the SSP possibly emerges from general auditory processing that favors sounds depicting an arch-shaped envelope, common amongst animal vocalizations".

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). For example, the SSP plays a role in cluster acquisition in both adults (Redford 2008) and children (Sprenger-Charolles & Siegel 1997), showing easier acquisition for those with a well-formed sonority profile. However, Davidson (2000) did not identify any variation of difficulty motivated by sonority in the production of non-native clusters by English speakers. In the present study, we propose to explore the possible SSP effects in the production of consonant clusters in word-onset position.

In this study, we investigate the potential SSP effects with an imitation task for #C1C2a or #C1əC2a structures, where C1C2 exhibits a sonority profile ranging from rising to falling. These sequences were produced by a native speaker of Tashlhiyt, a language allowing for word-initial #C1C2 sequences with rising, plateau, or falling profiles. They are presented to native Mandarin speakers, whose language prohibits any of these clusters. The aim is to investigate whether the SSP influences the production of onset clusters in native Mandarin speakers. Crucially, the imitation task includes the perception of the models to be imitated. The SSP effects on cluster perception have been observed in native Mandarin speakers (Zhao & Berent 2016; Chen et al. 2022): the repair #C1C2 > #C1əC2 was more frequent when C1C2 has an ill-formed sonority profile. In the current imitation task, the cluster productions

should also reflect what the participants have perceived. Thus, we anticipate perceptual repairs, which translate into a higher incidence of epenthetic vowels produced in consonant clusters with a more marked sonority profile, such as fall > plateau > rise. A purely perceptual account would predict similar schwas produced for the C1C2 and C1əC2 models, for example in terms of duration. Conversely, if the imitation task induces *production-specific* SSP effects, epenthetic vowels produced for C1C2 models should exhibit acoustic properties distinct from those of the fully produced vowels for C1əC2 models. We will thus examine the temporal and qualitative characteristics of the epenthetic vowels produced in the imitations. These data are needed to understand the intentional (targeted) or non-intentional (transitional) nature of the produced schwas and, possibly, to conclude whether or not schwas in the imitation data can be attributed solely to perception.

2. 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 C1C2a or C1əC2a, as detailed in Table 1. The C1C2 clusters exhibited rising, plateauing, or falling sonority profiles (referred to as k- or t-pivot for items with /k/ or /t/ in the C1 position for non-falling profiles or in the C2 position for falling profiles, respectively). C1əC2a items and C1C2a items differed solely in the presence of a schwa vowel between C1 and C2 in the former. All items were recorded eight times by the second author, a phonetician and native speaker of Tashlhiyt. The recorded items were scrutinized for the presence or absence of schwas within the C1C2 clusters. Two tokens of each item were chosen as models for the experiment. The models for C1C2a contained no vocalic material in the inter-consonantal position, while the models for C1əC2a included a schwa with a duration ranging from 42 to 100 ms ($\bar{x}$=71 ms, σ=18.4). 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.*

		C1C2a	C1əC2a
Rise	k-pivot	kla	kəla
	t-pivot	tla	təla
Plateau	k-pivot	kpa	kəpa
	t-pivot	tka	təka
Fall	k-pivot	lka	ləka
	t-pivot	lta	ləta

The experiment was conducted using the SpeechRecorder platform (Draxler & Jänsch 2004) in a quiet room, with an external sound card (Komplete Audio 6 MK2) and a headset microphone (AKG Pro Audio C544 L). For each trial, participants, seated in front of a computer, received a model stimulus once via headphones and were required to imitate the heard model without time pressure. No orthographic information was displayed on the computer screen during the sessions. 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 C1 and C2, 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 C1 release, and an interval after C1 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. Figure 1 shows examples of imitations for /kla/ from 3 different participants, in which the labeling of a schwa is distinct or unclear.

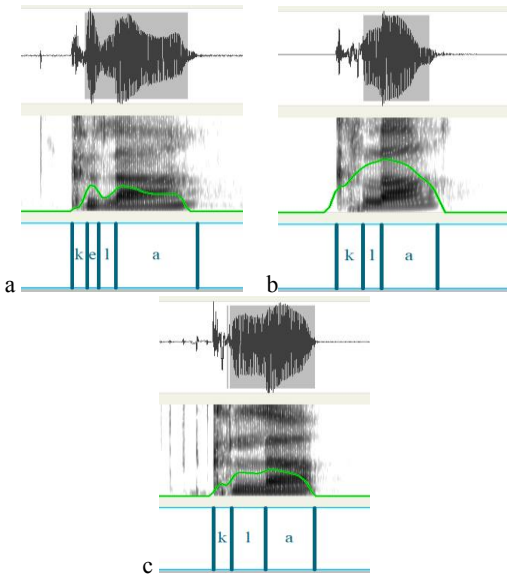


Figure 1: *Spectrograms and sound waves of 3 imitations for the item /kla/ for which the classification is clear (presence of schwa: a; absence of schwa: b) and ambiguous (c).*

3. Results

To examine the SSP effects, we compared the frequency of vocalic elements produced in different types of clusters. Results, presented in Figure 2, were analyzed using mixed effects logistic regression models, including Profile (rise, plateau, fall) and Condition (C1C2a, C1əC2a) as fixed effects. The random effects were random subject intercept and random subject slope on Condition. Condition was significant ($\chi^2(1)$=90.4, p<.001). C1əC2a yielded more vocalic elements than C1C2a (z=-7.5, p<.001). Within C1C2a, sonority falls yielded more vowels than plateaus (z=6.7, p<.001), and plateaus yielded more vowels than rises (z=5.2, p<.001). In C1əC2a, sonority falls from C1 to C2 also yielded more vowels than rises (z=3.9, p<.001) and plateaus (z=3.4, p<.01).

Concerning the quality of these vocalic elements, apart from a small number of vowels [i, u] (13%), there was a predominant tendency to produce a schwa between C1 and C2. To gain a better understanding of the nature of these inserted schwas, we additionally measured their acoustic characteristics. These include the relative duration of schwa, computed as the ratio between the absolute duration of schwa and the following vowel /a/, as well as the F1 and F2 frequencies at the midpoint of the schwa in C1C2 and C1əC2.

Figure 3 displays the results on relative duration. C1əC2a had significantly longer schwas than C1C2a (t(650)=-2.2, p<.01). In detail, for rising or plateau profiles, schwas for C1əC2a are significantly longer than for C1C2a (rising: t(165)=-2.8, p<.01, plateau: t(197)=-4.4, p<.001). This difference is not significant for the descending profile (t(284)=1.3, p=.02).

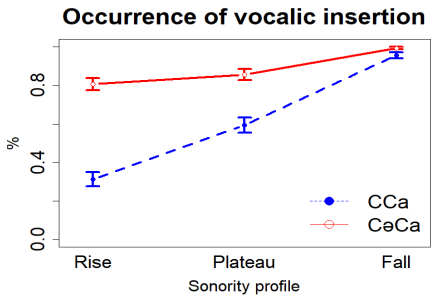


Figure 2: *Percentage of occurrence of vocalic elements according to sonority profile for C1C2a and C1əC2a items.*

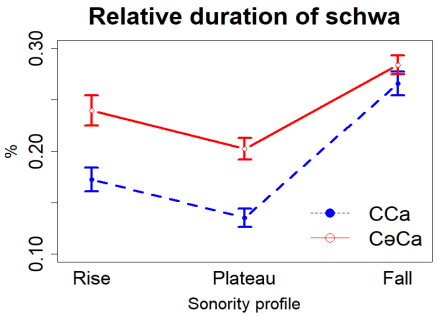


Figure 3: *Relative duration of schwas according to sonority profile for C1C2a and C1əC2a items.*

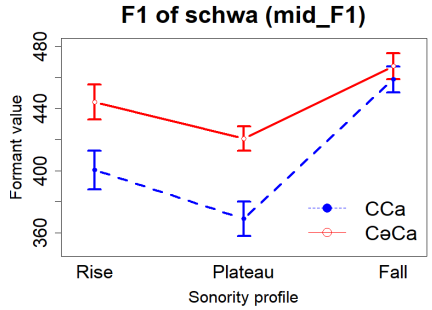


Figure 4: *F1 values at the midpoint of schwas according to sonority profile for C1C2a and C1əC2 items.*

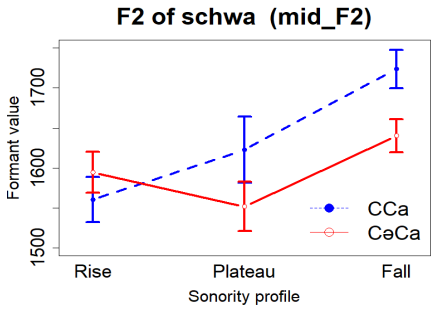


Figure 5: *F2 values at the midpoint of schwas according to sonority profile for C1C2a and C1əC2 items.*

The results on F1 and F2 at the midpoint of the schwas are respectively presented in Figure 4 and 5. The F1 values of the schwas produced for the items C1əC2a were significantly higher than those for C1C2a, both for the rising sonority profile (t(165)=-2.3, p<.05) and for the plateauing profile (t(197)=-3.9, p<0.001). There was no significant difference in F1 between schwas inserted in *ləka/ləta* and in *lka/lta* (t(284)=-0.7, p=.47). In terms of F2, the values were significantly higher for C1C2a than C1əC2a in the case of descending sonority (t(284)=2.6, p<0.01). However, this effect was not observed for the other two types of clusters (rise: t(165)=-0.8, p=.44; plateau: t(197)=1.4, p=.16). The effect Pivot significantly influenced F2 value for plateauing sonority profile (F(1, 198)=82.4, p<.001), with higher F2 values for t-pivot than k-pivot (t(197)=-9.1, p<0.001). However, this effect was not found for the other profiles (rise: t(165)=-0.9, p=.35, fall: t(284)=-0.06, p=.95).

4. Discussion and conclusion

In this study, we examined the SSP influence on the production of nonnative consonant clusters in word-initial position by Mandarin speakers. We found clear SSP effects in the onset clusters imitation task: the more marked the onset cluster, the more likely a vowel element was produced. Given the nature of the imitation task, it inherently involves aspects of both perception and production. 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; Chen et al. 2022), the data in Figure 2 likely reflects, in part, their perception of the model stimuli. As to the quality of inserted vowels, excluding the insertion of vowels [i, u] in certain clusters, whose presence suggests intentional epenthesis rather than purely transitional elements, we observe the presence of a large number of vocoids in C1C2a which closely resemble schwas. This raises the question of whether our imitation data solely represent perceptual repair of the model stimuli (i.e., intentionally inserted vowel) or whether they are modulated by the difficulty of producing consecutive consonants (i.e., non-intentional transitional vocoid). According to Articulatory Phonology (Browman & Goldstein 1992), transitional vocoids may result from insufficient (i.e., partial) gestural overlap between C1 and C2, due to certain articulatory constraints.

If the imitation data can be explained by perceptual effect alone, one would expect comparable schwas produced for both C1C2 and C1əC2 models. However, if an independent effect of SSP on production exists, some schwas in C1C2a should be the result of reduced gestural overlap between the two consonants, lacking their own articulatory target (i.e., difficulty of production). This would result in a set of acoustic characteristics that differ from those of the canonical schwa produced in C1əC2a (Davidson 2006; Gick & Wilson 2006; Ridouane & Fougeron 2011). In our data, schwa duration is shorter for C1C2 than C1əC2 (Fig. 3). This durational difference suggests that the schwas produced in C1C2 are more often unintended, transitional schwas compared to C1əC2. 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 C1C2 clusters. Analysis of schwa count, along with information on duration and F1, F2 of schwas in C1C2 and C1əC2 in Figures 3-5, demonstrates that schwa in C1C2 with falling sonority onsets closely resembles schwa in C1əC2 in terms of distribution, duration, and formant structure. The number of vocoids produced in /lk/ and /lt/ sequences is equivalent to the number of schwas produced in /lək/ and /lət/

sequences, with virtually identical durations and F1 values. For these highly marked onset clusters, it would be reasonable to conclude that the SSP effects are maximal and that the vowel element between C1 and C2 constitutes an epenthetic vowel resulting from perceptual repair, rather than merely a transitional element. The sonority profiles for the other two categories also align with SSP, showing the lowest proportion (~0.3) of produced schwas for /kl, tl/ and an intermediate proportion (~0.5) for /kp, tk/. In terms of their quality, schwas inserted for rising and plateau sonority profiles are shorter for C1C2 than C1əC2 models and differ in terms of F1. These data suggest that some of the schwas produced for C1C2 models are targetless and transitional, reflecting an SSP effect on the difficulty of production rather than perception. A shorter duration and a lower F1 could be attributed to the brief opening of a more closed vocal tract between two constrictions, as proposed by Flemming (2004). These results suggest that the production of C1C2 reflects not only an SSP effect in perception, leading to more targeted vowels for more marked clusters, but also an independent SSP effect in production, which manifests itself in the difficulty of production and is primarily observed in clusters with a non-falling profile.

In sum, we show that the SSP affects the production of consonant sequences in Mandarin, resulting in a higher occurrence of vowels within more marked clusters. This effect is not solely attributable to perceptual effects during the imitation task, but also indicates the existence of an SSP effect specific to the production. The perceptual SSP effect manifests itself in more targeted vowels within more marked C1C2 clusters, while the production-specific SSP effect resides particularly in the presence of transitional vowels in the production of rising and plateauing onset clusters. In order to test directly for a production-specific SSP effect, a comparison could be made with purely perceptual data: schwa-insertion should be more frequent in the imitation than in the perception task. Perceptual data are currently being analyzed with the same 20 subjects and the same items used in the imitation task. It is important to note that the assessment of schwa's status in the speech production of Chinese-speaking individuals is still preliminary and warrants further investigation, especially with more diverse datasets. One interesting aspect to explore is the effect of the articulatory position of neighboring consonants on F2, which may explain why F2 is higher in the falling profile and unchanged in the other profiles.

5. References

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., Lennertz, T., Jun, J., Moreno, M., & Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences*, 105(14), 5321–5325.

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.

Berent, I., Vaknin-Nusbaum, V., Balaban, E., & Galaburda, A. M. (2013). Phonological generalizations in dyslexia: The phonological grammar may not be impaired. *Cognitive neuropsychology*, 30(5), 285-310.

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(1), 35–59.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49(3-4), 155-180.

Chen, X., Ridouane, R., & Hallé, P. (2022). Perception des clusters selon leur profil de sonorité : le cas des auditeurs du mandarin confrontés à des clusters russes. *Proc. XXXIVe Journées d'Études sur la Parole -- JEP 2022*, 183-192.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. E. Beckman (Eds.), Laboratory Phonology. Cambridge: Cambridge University Press. 283–333.

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.

Davidson, L. (2006). Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics*, 34(1), 104-137.

Davidson, L., & Shaw, J. (2012). Sources of illusion in consonant cluster perception. *Journal of Phonetics*, 40(2), 234-248.

Draxler, C., & Jansch, K. (2004). SpeechRecorder - a universal platform independent multi-channel audio recording software. *International Conference on Language Resources and Evaluation*, 599-562.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, 25(6), 1568–1578.

Flemming, E. (2004). Contrast and perceptual distinctiveness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), Phonetically-Based Phonology. Cambridge: Cambridge University Press.

Gick, B., & Wilson, I. (2006). Excrescent schwa and vowel laxing: Cross-linguistic: responses to conflicting articulatory targets. In L. Goldstein, D. Whalen & C. Best, (Eds.), Laboratory Phonology. Berlin, New York: De Gruyter Mouton. 635-660.

Gómez, D., Berent, I., Benavides-Varela, S., Bion, R., Cattarossi, L., Nespor, M., & Mehler, J. (2014). Language universals at birth. *Proceedings of the National Academy of Sciences*, 111, 5837-5841.

Greenberg, J. H. (1978). Some generalizations concerning initial and final consonant clusters. In J. H. Greenberg (Eds.), Universals of Human Language. Stanford: Stanford University Press. 243-279.

Hallé, P., Segui, J., Frauenfelder, U., & Meunier, C. (1998). The processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception and Performance*, 24, 592-608.

Maïonchi-Pino, N., Taki, Y., Magnan, A., Yokoyama, S., Écalle, J., Takahashi, K., Hashizume, H., & Kawashima, R. (2015). Sonority-related markedness drives the misperception of unattested onset clusters in French listeners. *L'Année psychologique*, 115, 197-222.

Redford, M. A. (2008). Production constraints on learning novel onset phonotactics. *Cognition*, 107(3).

Ridouane, R., & Fougeron, C. (2011). Schwa elements in Tashlhiyt word-initial clusters. *Laboratory Phonology*, 2(2), 275-300.

Santolin, C., Crespo-Bojorque, P., Sebastian-Galles, N., & Toro, J. M. (2023). Sensitivity to the sonority sequencing principle in rats (Rattus norvegicus). *Scientific reports*, 13(1), 17036.

Sprenger-Charolles, L., & Siegel, L. (1997). A longitudinal study of the effects of syllabic structure on the development of reading and spelling skills in French. *Applied Psycholinguistics*, 18, 485-505.

Yin, R., Van de Weijer, J., & Round, E. (2023). Frequent violation of the sonority sequencing principle in hundreds of languages: how often and by which sequences? *Linguistic Typology*, 27(2), 381-403.

Zhao, X., & Berent, I. (2016). Universal restrictions on syllable structure: Evidence from Mandarin Chinese. *Journal of psycholinguistic research*, 45(4), 795–811.

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 de Paris, 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

Abstract

This study examines the impact of changes in two speaking styles –story reading vs. storytelling– on the spectral and temporal properties of long and short vowels in Jordanian Arabic. The transition from one register to another may generate temporal spectral modifications. This is why a particular interest has been paid to the behavior of long and short vowels in the context of these two types of variations. Ten speakers of Jordanian Arabic read and then narrated the same short story. Contrary to what was expected, spectral and temporal vowel properties were not influenced by the change in speaking style. These results indicate that in Jordanian Arabic, the transition from one register to the other had little impact on vowel quality and quantity. However, the conditions under scrutiny in this study may be too close to one another to enable such expected differences to emerge. Additional components of the currently collected corpus may be more appropriate to let differences between controlled and more spontaneous speech styles be revealed.

Keywords: speaking style, vowel length, Jordanian Arabic, spectral variation

1. Introduction

In continuous speech, the speaking style usually changes systematically depending on the situation that we experience. For example, in a classroom, we can read a text ("reading" style), talk to our teacher ("formal" style), and discuss with our classmates ("informal" style). This changing in 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 (Lindblom 1990; Farnetani and Recasens 2010). 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 (among others, DiCanio et al. (2015) in Arapaho, DiCanio et al. (2015) in Mixtec, Blaauw (1992) in Dutch, Bolotova (2003) in Russian, and Meunier and Espesser (2011) in French). 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[1]. 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.

Asymmetric influences were also noted between long and short vowels in speaking rate variation studies in several languages (Svastikula (1986) in Thai, Pind (1995) in Icelandic, and Hirata (2004) and Hirata and Tsukada (2009) in Japanese). According to these researches, the duration of long vowels is more lengthened than their short counterparts when the speaking rate slows down. The vocalic duration of long vowels is also more shortened than the duration of short vowels when the speaking rate accelerates. However, the impact of variation in rate on the vowel space seems to depend on the language. In Thai, spectral information of long and short vowels remains relatively stable (Svastikula 1986), unlike Japanese, the frequencies of short vowels are more influenced by the change of speaking rate than their corresponding short ones (Hirata and Tsukada 2009). In summary, the vowels, respectively long and short, react differently when the speaking rate or the speaking style is changed.

2. Research Question

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. Jordanian Arabic 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 Jordanian Arabic 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 (Al-Tamimi 2007; Abuoudeh 2018). According to the studies mentioned above, it is expected that reading a story can lead to longer vowel durations and larger spectral spaces than storytelling since the task of reading would correspond to hyper-articulated speech while the task of storytelling would be closer to a more hypo-articulated speech style. Furthermore, this influence could be asymmetrical between short and long vowels.

[1] An endangered Algonquian language spoken in the State of Wyoming in the United States of America. This language have phonological length opposition and contains 4 long and 4 short vowels.

3. Methods

3.1. Speakers

To answer the problem of this study, 10 Jordanian speakers (5 females and 5 males) participated voluntarily in a speech production experiment. The participants were all undergraduate students at Al-Hussein bin Talal University in Ma'an, in the south of Jordan and were aged between 18 and 22 at the time of the recording. They are from Amman and Zarqa, cities located in the Central region of Jordan. The speakers have declared that they do not have any speech disorder.

3.2. Stimulus

The stimulus for this experience consists of the story of "Little Red Riding Hood" written in the Arabic alphabet in a version of Jordanian Arabic written by the first author. It should be noted that this story is popular in Jordan, and all of the registered participants declared that they knew it. The choice of a well-known and popular story is intended to facilitate the task of storytelling[2].

3.3. Procedure

First, speakers were asked to read the story of 'Little Red Riding Hood' from a text that was displayed on a computer screen. Subsequently, they were asked to tell the same story, without reading it. Before recording the storytelling task, the speakers could – if they felt it necessary – reread the story silently to prepare their narration. Before the experiment began, participants were instructed to read and retell the story in their dialect and not in Classical Arabic.

The experiment took place in a quiet room at the Faculty of Letters of Al-Hussein bin Talal University. The equipment used for the recordings is a Sennheiser e835 microphone connected to a Tascam DR-100. The sound files were sampled at 44100 Hz on 32 bits in monophonic mode. The recordings of the two tasks (reading and storytelling) were first transcribed and transliterated with the new Arabic transliteration system (ATR convention) and then segmented by forced alignment using the 'Arabic WebMAUS Basic' service (Kisler, Reichel, and Schiel 2017; Al-Tamimi et al. 2022).

The results of the forced alignment were subsequently corrected by hand using the Praat software (Boersma and Weenink 2022). The duration of the segments, the frequency of the formants (F1, F2, F3), and the f0 were automatically extracted by a Praat script. The *Burg* extraction algorithm (LPC analysis by autocorrelation) was used with an analysis window of 0.025 s and a step of 0.01 s. The formant extraction thresholds were adapted to the sex of the speaker (5000 Hz maximum for men and 5500 Hz maximum for women). The extracted data was then saved in a *.csv* file. For this study, the duration and frequencies of the F1 and F2 formants of vowels were analyzed. The frequencies of F1 and F2 of all speakers were normalized using the Lobanov method in order to limit inter-speaker variation (Lobanov 1971)[3]. Data analyses were performed using the R program (R Core Team 2021).

[2]The data from this study are part of a larger database that is currently under construction on Jordanian Arabic ("Speech Database Jordanian Arabic Dialects - SDJAD" project), which will consist of over 100 participants from different regions of Jordan.

[3]Normalization was carried out using the function 'normLobanov' from the library 'phonR' (McCloy 2016).

3.4. Statistical analysis

The relationships between each of the studied dependent variables ("Vowel duration", "F1", and "F2") and the fixed effects ("Vowel" and "Task") were evaluated by linear mixed models with the function 'lmer' from the library 'lme4' (Bates et al. 2015). The intercept for speakers was also included in the models as a random effect. Additionally, per-speaker random slopes were included for each fixed effect, corresponding to the inter-speaker variability in the effect of each fixed factor on the dependent variables to avoid a high rate of Type I error. The *p-values* were obtained by Satterthwaite approximations using the 'anova' function from the 'lmerTest' library (Alexandra Kuznetsova 2017). These analyses were followed by *post hoc* Tukey tests using the 'glht' function of the 'multcomp' library (Hothorn, Bretz, and Westfall 2008).

4. 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. Furthermore, it should be noted that

Vowel	Task	
	reading	telling
i	1120	942
iː	393	360
a	1664	1211
aː	1185	871
u	81	155
uː	188	182
eː	278	180
oː	63	91

Table 1: Number of realisations of each vowel in each speaking style.

short vowels – except /u/ – are overall more frequent than long vowels in the present data, regardless of the speaking style, with a total of 5173 short vowels compared to a total of 3791 long vowels.

4.1. Duration

Descriptive analyses indicate that the two studied speaking styles have a low impact on vowel durations (Figure 1). Mean durations of short vowels remain relatively stable in both speaking styles. As for those of long vowels, the /iː, oː/ are slightly longer in reading than in storytelling. The vowels /aː, uː/, on the contrary, are longer in narration than in reading, particularly the duration of /uː/. The duration of /eː/ remains relatively unchanged in both styles. The observations from the descriptive analyses were confirmed by linear mixed analyses that show no significant difference between the task of reading and the task of storytelling for the duration ($F_{(1,7)} = 0.30, p = .587$). In addition, *post hoc* analyses (Tukey) point out that the duration of vowels is not significantly different depending on the speaking style except for /iː, uː/. These results also reveal that the temporal relationship between long and short vowels in Jordanian Arabic is not influenced by changing speaking style from reading to storytelling.

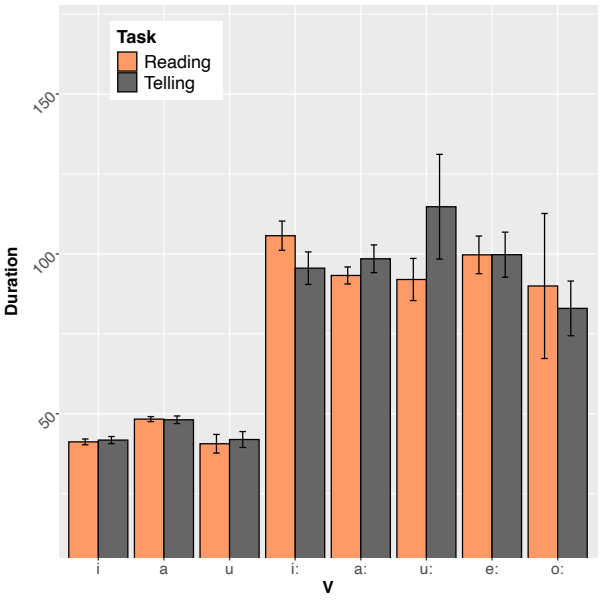


Figure 1: Means of vowel durations in the two speaking style conditions (in ms, the error bars represent the Confidence Interval at 95%).

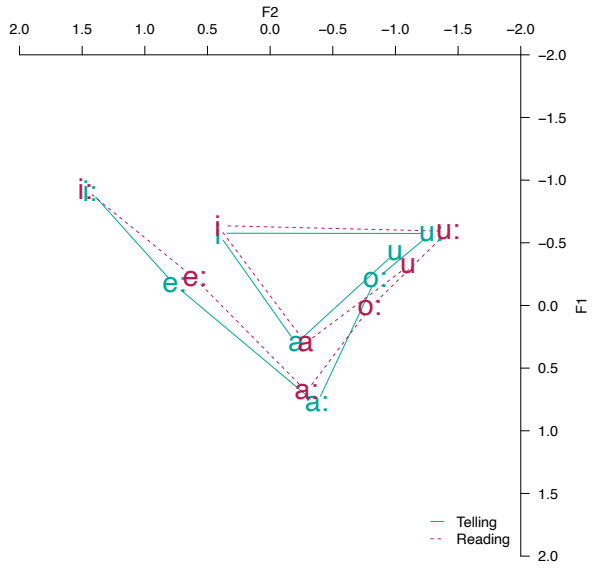


Figure 2: Vocalic space of the eight vowels (in Lobanov) in both speaking style conditions.

4.2. Spectral space

The examination of the vowel space highlights also that the two speaking styles have little influence on spectral information (Figure 2). Indeed, long and short vowels occupy very close positions in both speaking styles on the F1-F2 space. These observations were confirmed by linear mixed analyses, which showed no significant difference between the reading task and the storytelling task for the frequencies of F1 ($F_{(1,7)} = 0.48, p = .494$), and of F2 ($F_{(1,7)} = 0.0001, p = .99$). The *post hoc* analyses (Tukey) confirm also that the frequencies of F1 and F2 observed for all vowels do not significantly change when speaking style changes. Furthermore, these results reveal that the spectral relationship between long and short vowels in Jordanian Arabic is not influenced by the change in speaking style from reading to storytelling.

5. Discussion and conclusion

This study aimed at evaluating the impact of changing speaking style on vowel opposition in Jordanian Arabic. According to the results of this study, this change has very little influence on the spectral and temporal information of long and short vowels. The vowel quality showed no significant difference between the two speaking styles for all vowels. As for the quantity, only two vowels out of eight revealed a significant difference depending on the style (/uː/ and /iː/), including one in an unexpected direction. Indeed, the vowel /uː/ – but also slightly the /aː/ with no significant effect – attests to a lengthening of its duration in storytelling rather than reading. This observation could be due to more hesitation or reflection in the storytelling task than in the reading task.

These findings are not in agreement with previous studies mentioned above. As a reminder, these studies described that the transition from formal to spontaneous speech leads to spectral and temporal variations that can be asymmetric between long and short vowels. The findings of this research could be explained by the fact that these two speaking styles have potentially limited effects on temporal differences in a language that contains a phonemic length opposition. In other words, the absence of the reading *vs.* storytelling distinction – in the case of this study – could be due to the proximity of the two styles compared to the styles investigated in the studies mentioned above. For example, DiCanio et al. (2015) – but also DiCanio and Whalen (2015) – describe that their condition "elicitation" is a repeated pronunciation of isolated words and that the "spontaneous" speech is taken from telling a personal story. It is potentially significantly more discriminating in speaking style terms than reading *vs.* storytelling of the same story, such as that which we compare in the present study.

In addition, the importance of duration separation between long and short vowels in Jordanian Arabic could reduce the temporal impact and, therefore, the variations associated with spectral space in these two types of speaking styles. Another factor for this absence of style effect is that Jordanian Arabic speakers are not used to reading stories in the Jordanian Arabic dialect since they mainly read stories in classical Arabic. This may explain why their reading style resembles closely to their storytelling style. During the recordings, a hesitation, even a reflection, was observed with some speakers in both speaking styles. Finally, studying other tasks of the SDJAD project (such as words produced in isolation, conversational speech, and image description) that are in progress could be enriching to evaluate these different assumptions.

6. Acknowledgements

This research was funded by Al-Hussein Bin Talal University, grant number (85/2022).

7. References

Abuoudeh, Mohammad (2018). “De l’impact des variations temporelles sur les transitions formantiques”. PhD thesis. Université de Nantes.

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), © European Language Resources Association (ELRA), Licensed under CC-BY-NC-4.0*. Marseille, 20-25 June 2022, pp. 7269–7276.

Al-Tamimi, Jalal-Eddin (2007). "Indices dynamiques et perception des voyelles : Étude translinguistique en arabe dialectal et en français". Thèse de doctorat. Université Louis Lumière - Lyon 2, p. 580.

Alexandra Kuznetsova Per B. Brockhoff, Rune H. B. Christensen (2017). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 82.13, pp. 1–26. DOI: https://doi.org/10.18637/jss.v082.i13.

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: https://doi.org/10.18637/jss.v067.i01.

Blaauw, Eleonora (1992). "Phonetic differences between read and spontaneous speech". In: *II International Conference on Spoken Language Processing ICSLP*.

Boersma, Paul and David Weenink (2022). *Praat: doing phonetics by computer [Computer program]. [Computer program]*. Version 6.2.09. URL: http://www.praat.org/.

Bolotova, Olga (2003). "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)*.

DiCanio, Christian and D.H. Whalen (2015). "The interaction of vowel length and speech style in an Arapaho speech corpus". In: *The 18th International Congress of the Phonetic Sciences*.

DiCanio, Christiani, Hosung Nam, Jonathan D. Amith, Rey Castillo García, and D. H. Whalen (2015). "Vowel variability inelicited versus spontaneous speech: Evidence from Mixtec". In: *Journal of Phonetics* 48, pp. 45–59.

Farnetani, Edda and Daniel Recasens (2010). "Coarticulation and connected speech processes". In: *The Handbook of Phonetic Sciences*. Ed. by W. J. Hardcastle, J. Laver, and F. E. Gibbon. Second. Wiley-Blackwell, pp. 316–352.

Hirata, Yukari (2004). "Effects of speaking rate on the vowel length distinction in Japanese". In: *Journal of Phonetics* 32, pp. 565–589.

Hirata, Yukari and Kimiko Tsukada (2009). "Effects of speaking rate and vowel length on formant frequency displacement in Japanese". In: *Phonetica* 66, pp. 129–149.

Hothorn, Torsten, Frank Bretz, and Peter Westfall (2008). "Simultaneous Inference in General Parametric Models". In: *Biometrical Journal* 50.3, pp. 346–363.

Kisler, Thomas, Uwe Reichel, and Florian Schiel (2017). "Multilingual Processing of Speech via Web Services". In: *Comput. Speech Lang.* 45.C, 326–347. DOI: 10.1016/j.csl.2017.01.005. URL: https://doi.org/10.1016/j.csl.2017.01.005.

Lindblom, Björn (1990). "Explaining phonetic variation : A sketch of H&H Theory". In: *Speech production and speech modelling*. Ed. by W.J. Hardcastle and A. Marchal. Kluwer Academic Publishers, pp. 403–439.

Lindblom, Björn and Rolf Lindgren (1985). "Speaker-listener interaction and phonetic variation". In: *Phonetic Experimental Research at the Institute of Linguistics University of Stockholm-PERILUS* 4, pp. 77–85.

Lobanov, B. M. (Feb. 1971). "Classification of Russian Vowels Spoken by Different Speakers". In: *The Journal of the Acoustical Society of America* 49.2B, pp. 606–608. DOI: 10.1121/1.1912396. eprint: https://pubs.aip.org/asa/jasa/article-pdf/49/2B/606/18770434/606_1_online.pdf. URL: https://doi.org/10.1121/1.1912396.

McCloy, Daniel R. (2016). *Normalizing and plotting vowels with phonR 1.0.7*. URL: http://drammock.github.io/phonR/.

Meunier, Christine and Robert Espesser (2011). "Vowel reduction in conversational speech in French: The role of lexical factors". In: *Journal of Phonetics* 39.3, pp. 271–278. DOI: https://doi.org/10.1016/j.wocn.2010.11.008. URL: https://www.sciencedirect.com/science/article/pii/S0095447010000951.

Pind, Jörgen (1995). "Speaking rate, voice-onset time, and quantity: The search for higher-order invariants for two Icelandic speech cues". In: *Perception & Psychophysics* 57.3, pp. 291–304.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Svastikula, M. L. Katyanee (1986). "A perceptual and acoustic study of the effects of speech rate on distinctive vowel length in Thai". PhD thesis. The University of Connecticut, p. 110.

Timing of acceleration peaks and acceleration changes

Malin Svensson Lundmark

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

Abstract

Articulators accelerate and decelerate continuously during speech. Previous research reveals a structured pattern of deceleration peaks aligning with segment onset and acceleration peaks with segment offset. This study reports on previous EMA findings on 18 speakers while also sets out to explain why the deceleration peak lags behind at onset boundary, as previously reported. A qualitative analysis is made on acceleration changes (jerk) of constrictions by one speaker. Preliminary results reveal significant differences between the kinematics of segment onset at constriction closure, which align to the jerk of the deceleration phase, and segment offset at the release, which coincide with the acceleration peak.

Keywords: speech production, acceleration peak, jerk, EMA, segmental articulation

1. Introduction

Acoustic segment transitions have been shown to correspond to acceleration peaks of primary active articulators (Svensson Lundmark 2023). Specifically, for any speech posture of an active articulator we find a deceleration peak at the start of the speech posture and an acceleration peak at the end of the posture, seemingly dividing the speech postures from the fast intervals of the movements to and from the postures (Svensson Lundmark and 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, Pendrill, and Reistad 2016). In **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 a static speech posture, changes direction mid-posture (dashed vertical line in **Figure 1**), and then moves rapidly away again (an acceleration peak, followed by a velocity peak). In speech production, an active articulator constantly moves in and out of speech postures just like this, delimited by fast intervals, and acceleration peaks. This pattern can be seen in all articulators, even when it's not crucial for making a constriction, as described in the DASA approach (Descriptive Approach to Segmental Articulations) in Svensson Lundmark and Erickson (2024).

The deceleration and acceleration peaks coincide with the edges of an articulatory posture and in extension with the segment boundaries (Svensson Lundmark 2023), as visualized by the 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

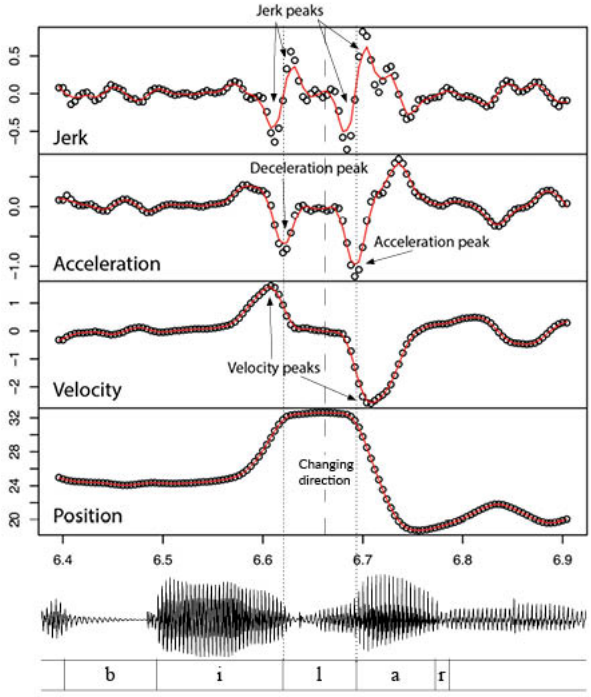


Figure 1: *From bottom up: speech signal, vertical tongue tip movement (position), and velocity, acceleration and jerk signals (smoothed signals in red), while producing /l/ in <bilar>.*

and place of articulation (Svensson Lundmark 2021; Svensson Lundmark 2023; Svensson Lundmark and Frid 2023; Svensson Lundmark and Erickson 2024). However, there appears to be a small but consistent lag between the segment boundary and the de/acceleration peak. This pattern has been most prominent at the onset of the segment where the deceleration peak lags behind the acoustic boundary by approximately 10 ms (Svensson Lundmark 2023; Svensson Lundmark and Erickson 2024), which might indicate a coordination at segment onset with something other than the deceleration peak.

On the top row in **Figure 1** you find jerk (third derivate to position). A jerk peak occurs when acceleration changes the most (Eager, Pendrill, and Reistad 2016). Hence, rapid position changes of any object, be it a rollercoaster or the sudden braking of a car, are often referred to as "jerky" movements. When we speak, we produce jerk too, as visualized in **Figure 1**. Jerk peaks appear on either side of the de/acceleration peaks as acceleration changes both before and after its maximal value.

The previously reported short but consistent lag between a deceleration peak and C1 and C2 segment onset boundary may

be because of a coordination with the acceleration change (=jerk peak) rather than the acceleration peak itself. This study examines this possibility by reporting on some previous findings on the relationship between de/acceleration peak and acoustic segment boundary, and in addition present preliminary results on the timing of acceleration changes (jerk peaks) of lower lip and tongue tip.

2. 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. For detailed information on the procedure, see Svensson Lundmark (2023) and Svensson Lundmark and Erickson (2024). EMA position data was collected from a number of articulators. Here is reported on sensors on lips, tongue tip, tongue dorsum, and lower incisors (lower jaw). Post-processing of signals was done in Carstens software, and in R (R Core Team 2021), where specifically calculation and articulatory analyses were performed. Acoustic segmentation was done by the author in Praat (Boersma and Weenink 2009). An inter-annotator agreement (IAA) was also performed showing that 93,4% of the segment boundaries differed by less than 10 ms. See more details on the IAA in Svensson Lundmark (2023).

Lip aperture (LA) was calculated in R using the three-dimensional Euclidian distance between the sensors on the upper and the lower lip. While LA is a calculated three-dimensional distance (x, y, z) between two sensors, measures on the tongue tip (TT), tongue dorsum (TD) and the jaw (JW) are instead based on the two-dimensional positions of each sensor (x, y), i.e. tangential velocity. 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, with a low span (0.1) for the smoothing not to cause any distortion. The value was determined by visually inspecting the result. The acceleration signal is simplified by using loess and only smoothed for the purpose of collecting the data, but it should be noted that the signals have already been filtered during the data processing with the Carstens software.

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 were calculated by measuring the time lag to the expected corresponding acoustic segment boundary. Example of time lag measurements on LA, TT and JW can be seen in **Figure 2**. Here the target word is /man:a/ where the lips are the primary articulator at syllable onset and TT is the primary articulator at syllable coda position.

As a statistical tool to evaluate the time lags, linear mixed effects models (LMM) were used and run in R. For details on statistical analysis see the papers where the results were originally published (Svensson Lundmark 2023; Svensson Lundmark and Erickson 2024).

2.1. Acceleration changes (jerk)

To investigate the possibility of segment onset aligning with the acceleration change rather than the deceleration peak, this study

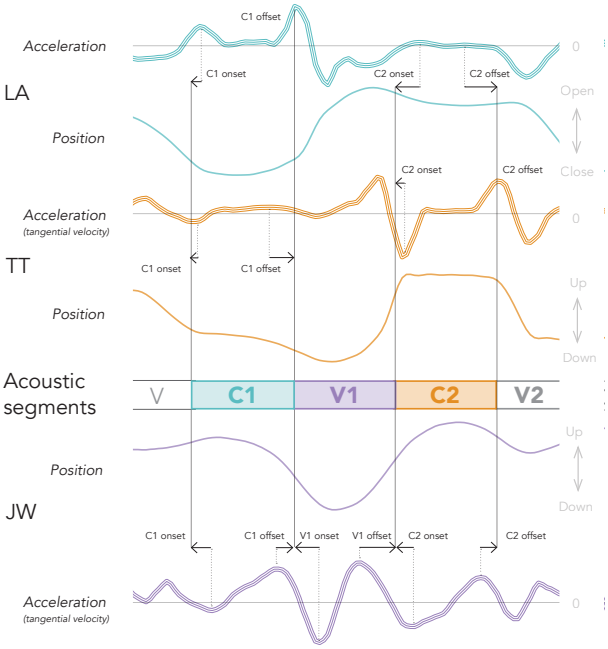


Figure 2: *Example of time lag measurements on LA, TT and JW (TD not represented in this figure). Colored solid curves are position signals. Triple lines represent the acceleration signals. From a female speaker and the target word /man:a/. Time lags are measured between a segment onset and a deceleration peak, and between a segmental offset and an acceleration peak. Reprinted from Svensson Lundmark and Erickson (2024).*

makes a qualitative assessment of one female speaker and her production of /b, m, p, n/. For the preliminary analyses of acceleration changes, jerk is used, which is derived by computing the third-order differences of the position data. The jerk signal was smoothed similarly as the acceleration signal, using the R-function loess with a span of 0.1. However, the jerk signal was solely based on vertical positions of lower lip (LL) and TT. As the preliminary results on jerk peaks include only qualitative analysis and visual presentations of the constrictions from only one speaker, no statistical analysis was performed.

3. 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) and Svensson Lundmark and Erickson (2024). The table only includes results from measures on primary articulators, such as LA in /m/, and TT in /n/. For a constriction with lips or TT 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 (10-12 ms).

However, the deceleration and acceleration landmarks of TD are not aligned to V1 onset and V1 offset (**Table 1**). Instead we find long and varied time lags indicating a much shorter tongue body speech posture than vowel segment. Note that the acceleration peak of the primary articulator at C1 offset, and the deceleration peak of the primary articulator at C2, determine the acoustic vowel segment duration, regardless of whether it is LA or TT.

Table 1: *Results reported from Svensson Lundmark (2023) and Svensson Lundmark and Erickson (2024) on average time lags in ms (stdv in italics) of de/acceleration peaks to acoustic segment boundaries. Only time lags on primary articulators are included (i.e. LA for /m/, TT for /n/, TD for /a/), with the exception of JW. A positive time lag means the de/acceleration peak precedes the segment boundary; a negative that the de/acceleration peak follows the segment boundary.*

	C1ons	C1off	V1ons	V1off	C2ons	C2off
	Dec	Acc	Dec	Acc	Dec	Acc
LA	-11 *5*	-4 *10*	-	-	-10 *4*	2 *8*
TT	-12 *10*	-5 *8*	-	-	-10 *6*	2 *7*
TD	-	-	-35 *20*	25 *20*	-	-
JW	-20 *15*	15 *15*	-40 *15*	45 *20*	-20 *12*	17 *15*

The lower jaw displays overall long and varied time lags, as seen in **Table 1**. The deceleration peak always follows the segment onset by approximately 20 ms, while the acceleration peak always precedes the segment offset by approximately 15 ms (the lags are larger at V1 onset and V1 offset) Timing of the JW de/acceleration peaks tell us that the jaw speech postures are shorter than the postures of the other articulators (LA, TT and TD).

3.1. Preliminary results on acceleration changes (jerk)

Results include EMA positions and calculations from one Swedish speaker. **Figure 3** shows the TT vertical position and velocity, and above that acceleration and jerk signals in the word <nanna> (a personal name), where stress is on the first syllable (bold letters in the bottom of **Figure 3**). As marked in the figure, the deceleration peak lags behind both C1 and C2 segment onset. Instead the segment boundary seems better aligned to the jerk peak. This jerk peak is the start of the deceleration phase towards the constriction, as it coincide with the velocity peak of the closure (the peaks on the dotted curve). For C1 and C2 segment offset, alignment with the boundary is seen for both the acceleration peak and the preceding jerk peak. However, this jerk peak does not coincide with a velocity peak as it occurs at the end of the much slower posture.

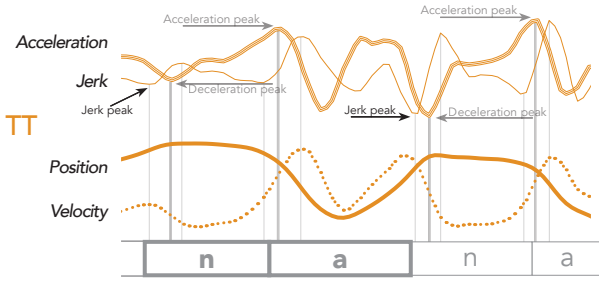


Figure 3: *The target word /nan:a/ (stress on the first syllable) with vertical TT movement (position, solid line), and velocity (dotted line), acceleration (triple lines) and jerk signals (thin solid line). Vertical triple lines mark deceleration and acceleration peaks, vertical thin solid lines mark jerk peaks.*

If we turn to lower lip movements in /m/ we see a similar pattern in the word <mamma> (mother in Swedish) (**Figure 4**). The deceleration peaks at C1 and C2 onset lag behind the segment boundary, and instead the jerk peak (a velocity peak during the closure) is a much better fit for timing between the acoustic and articulatory landmarks. The acceleration peak at C1 and C2 segment offset is nicely aligned with the annotated acoustic segment boundary.

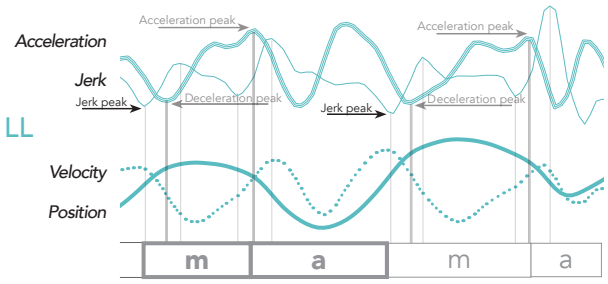


Figure 4: *The target word /bi:lar/ (stress on the first syllable) with vertical LL movement (position, solid line), and velocity (dotted line), acceleration (triple lines) and jerk signals (thin solid line). Vertical triple lines mark deceleration and acceleration peaks, vertical thin solid lines mark jerk peaks.*

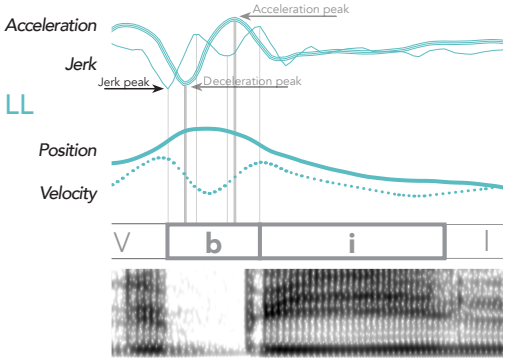


Figure 5: *The target word /bi:lar/ (stress on the first syllable) with vertical LL movement (position, solid line), and velocity (dotted line), acceleration (triple lines) and jerk signals (thin solid line). Vertical triple lines mark deceleration and acceleration peaks, vertical thin solid lines mark jerk peaks.*

Peak velocities of LL and TT at offset occur well within the vowel segment (**Figure 3** and **Figure 4**). However, this pattern depends on the constriction. If we turn to constrictions with built up intra-oral pressure, as in /b/ in <bilar> (cars) and /p/ in <pappa> (father), the velocity peak at the release of LL co-occurs with the vowel segment onset (**Figure 5** and **Figure 6**). The acceleration peak is instead aligned with the release burst, as this marks the time the speech posture ends. As before, the deceleration peak of the closure lags behind segment onset boundary of both /b/ and /p/, and the jerk at the velocity peak is timed with the segment onset boundary.

4. Discussion and conclusion

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, as summarized in **Figure 7**. The speech postures of the consonantal articulators, delimited by deceleration and acceleration peaks, shape the

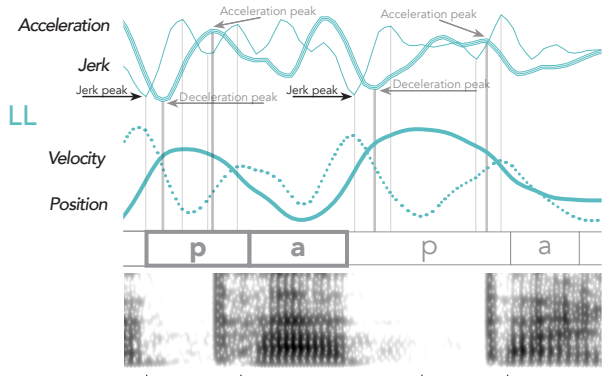


Figure 6: *The target word /pap:a/ (stress on the first syllable) with vertical LL movement (position, solid line), and velocity (dotted line), acceleration (triple lines) and jerk signals (thin solid line). Vertical triple lines mark deceleration and acceleration peaks, vertical thin solid lines mark jerk peaks.*

segment durations, while the speech postures of the lower jaw appear to be much shorter than the postures of any other articulator (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 is not captured sufficiently by one EMA sensor.

Results also include a first attempt of explaining why deceleration peaks lag behind the onset segment boundaries. Basically, it may be because of a coordination with the acceleration change (the jerk peak) rather than the deceleration peak. This acceleration change begins at the fastest moment - velocity peak - where the movement starts to decelerates. After the deceleration peak the movement is still, hence the speech posture. At mid speech posture the articulator starts to accelerate again and reaches its acceleration peak at the segment offset, or at release burst, depending on the constriction. The different nature of deceleration and acceleration of a movement is transferred to the different nature of acoustic segment transitions; one is the result of the start of a deceleration phase of the closure, and the other is the result of acceleration peak of the release, i.e. maximum added force. That closure and release are inherently different is a well-known phenomenon, but one that is rarely investigated in detail, not least with a focus on acceleration and jerk.

In previous research we have reported that the deceleration and acceleration peaks align with segmental boundaries (Svensson Lundmark 2023; Svensson Lundmark and Erickson 2024). This may need to be redefined as this study reports that consonants consist of an onset of closure (a jerk that starts off the deceleration) and an offset of closure (an acceleration peak). Further research will determine whether this pattern is speaker- and language-independent.

Thorough investigation of the acoustic-articulatory relationship of the closure and the release of constrictions may give insights into how planning of speech units are related to the articulatory movement patterns. These preliminary findings on acceleration changes hypothesize two separate movements in speech posture production. Such a structure work in connection with theories on speech production that account for articulatory movements reaching a target, as well as those focused on the initiation of an articulatory movement.

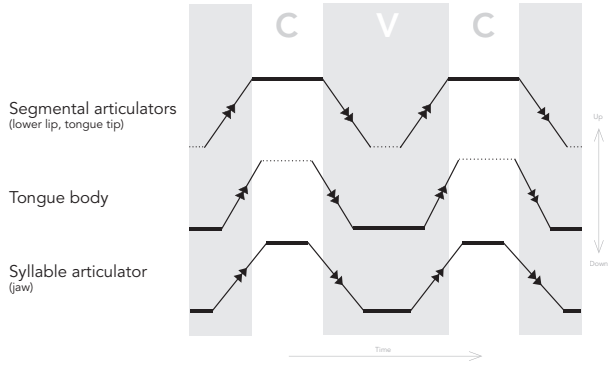


Figure 7: *A schematized figure of articulatory intervals as divided by de/acceleration peaks, based on results in Svensson Lundmark (2023) and Svensson Lundmark and Erickson (2024). Grey/white areas mark acoustic segment duration. Vertical solid lines are speech postures. Slanted lines with arrows are fast intervals. Reprinted from Svensson Lundmark and Erickson (2024).*

5. Acknowledgements

This work was supported by an International Postdoc grant from the Swedish Research Council (Grant No. 2021-00334).

6. References

Boersma, Paul and David Weenink (2009). *Praat: doing phonetics by computer (Version 5.1.13)*. URL: http://www.praat.org.

Eager, David, Ann-Marie Pendrill, and Nina Reistad (2016). "Beyond velocity and acceleration: jerk, snap and higher derivatives". In: *European Journal of Physics* 37.6. DOI: 10.1088/0143-0807/37/6/065008.

Fujimura, Osamu (1961). "Bilabial stop and nasal consonants: A motion picture study and its acoustical implications". In: *Journal of Speech and Hearing Research* 4.3, pp. 233–247.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Svensson Lundmark, Malin (2021). "Evidence of segmental articulations: Acceleration determines vowel segment duration in Swedish Word Accents". In: *Proc. 1st International Conference on Tone and Intonation (TAI)*, pp. 156–160. DOI: 10.21437/TAI.2021-32.

— (2023). "Rapid movements at segment boundaries". In: *The Journal of the Acoustical Society of America* 153.3, pp. 1452–1467. DOI: 10.1121/10.0017362.

Svensson Lundmark, Malin and Donna Erickson (2024). "Segmental and syllabic articulations: a descriptive approach". In: *Journal of Speech, Language, and Hearing Research*. DOI: https://doi.org/10.1044/2024_JSLHR-23-00092.

Svensson Lundmark, Malin and Johan Frid (2023). "Segmental articulations across prosodic levels". In: Niebuhr, Oliver and Malin Svensson Lundmark. *Proceedings of the 13th International Conference of Nordic Prosody*. Sciendo, pp. 255–261. DOI: 10.2478/9788366675728-023.

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], Jacques Felblinger[3,4], Yves Laprie[5]

[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

Abstract

This communication focuses on the production of ejectives and implosives in Human Beatboxing (HBB). To investigate ejectives and implosives (i.e. glottalic mechanisms), real-time Magnetic Resonance Imaging data was collected on one professional subject. MRI recordings were analyzed via a functional Principal Component Analysis of vocal tract contours. The findings show that glottalic mechanisms for this subject are produced with tongue root and velo-pharyngeal maneuvers to change the volume of the pharyngeal cavity rather than laryngeal raising or lowering. We think ejectives and implosives might better be described as obstruent consonants produced with pharyngeal compression or expansion and a closed glottis at least in Beatboxing.

Keywords: Human Beatboxing, rt-MRI, Ejectives, Implosives

1. Introduction

This communication focuses on the production of ejectives and implosives in Human Beatboxing (HBB). Human Beatboxing is a vocal technique where musicians imitate musical sonorities, such as instruments or electronic music, with their vocal tract. Human Beatboxing is a novel paradigm to study sound production mechanisms in the vocal tract among a population of highly trained subjects. To investigate ejectives and implosives, 2D real-time magnetic resonance imaging (rt-MRI) data was collected on one professional subject. The goal is to know whether ejectives and implosives in HBB are produced similarly or differently to glottalic mechanisms of the world's languages.

1.1. Ejectives and implosives of the world's languages

Most studies focus on the acoustic properties such as Voice Onset Time (VOT), burst intensity or F0 perturbations induced by the so-called glottalic mechanism on the following vowel. Little attention has been given to the glottalic mechanism itself.

Ejectives are produced by decreasing the volume between the closed glottis and the oral constriction, resulting in high positive pressure. Conversely, implosives are produced by increasing the volume in the supralaryngeal region between the closed glottis and the oral constriction, resulting in low negative pressure (Catford 1977; Ladefoged 1971). Volume and pressure variations during the so called glottalic stops are attributed to laryngeal raising (ejectives) and lowering (implosives).

Kingston (1985) attempted to run different aerodynamic models of stops in Tigrinya. The simulated data showed that laryngeal raising failed to increase sufficiently IOP to reach the expected pressure values of Tigrinya ejective stops. Increased contractions of the supraglottal cavity resulted in the expected values of pressure for ejectives. He attributes the additional contractions to tongue root retraction and/or vocal tract walls stiffening.

An MRI study of Oh and Lee (2018) found higher larynx position for ejectives compared to implosives in Hausa. Though, no comparison was made with pulmonic stops. Another MRI study of Sulaberidze et al. (2023) showed higher laryngeal position for ejectives stops compared to pulmonic stops in Georgian. They express doubts on sufficient laryngeal movement to increase IOP for ejectives. An interesting descriptive study of Hermes et al. (2016) using MRI reported tongue root retraction and advancement of the posterior wall of the pharynx to decrease the size of the supralaryngeal cavity during the production of ejectives in Tigrinya.

The role of laryngeal lowering during the production of implosives has also been questioned by Demolin (1995) and Demolin, Ngonga-Ke-Mbembe, and Soquet (2002). Based on MRI, they showed pharyngeal expansion during the production of implosives. The expansion results from laryngeal lowering and tongue root advancement. They question the relationship between tongue root advancement and laryngeal lowering.

Tongue root maneuvers might be involved in the production of both ejectives and implosives. Though, it is not clear how it relates to laryngeal height.

1.2. Ejectives and implosives in Human Beatboxing

The MRI recordings of professional beatboxer in the study of Proctor et al. (2013) reported systematic laryngeal lowering followed by an oral closure, a glottal closure and laryngeal raising. Similar characteristics have also been reported in other MRI studies (Patil et al. 2017; Dehais Underdown et al. 2023). Results on laryngoscopic investigations show pharyngo-laryngeal dilatation followed by pharyngo-laryngeal contraction. The pharyngeal compression is provoked by laryngeal raising and an epiglottal retraction while the glottis remains closed (De Torcy et al. 2014; Fabre 2018; Dehais-Underdown et al. 2021).

In Dehais Underdown et al. (2023), the authors reported systematic involvement of the tongue root in both ejectives and implosives but they did not quantify their observations. Similar observations were made in De Torcy et al. (2014) and Fabre (2018) who observed retraction of the epiglottis in their laryngoscopic analysis. Whether tongue root movements act independently of laryngeal raising or lowering in the production of

beatboxed ejectives and implosives remains to be verified.

Concerning aerodynamic studies of beatboxing, in his master's thesis, Fabre (2018) reports IOP values ranging from 20hPa up to 80hPa for 1 professional beatboxer. In Dehais-Underdown et al. (2021), the IOP reported for 5 subjects producing beatboxing ejectives reached between 40hPa and 100hPa. They also reported a voiceless bilabial implosive [ɓ̥] with high negative pressure reaching about −95hPa.

Along with Kingston's (reasonable) doubts on the sufficient laryngeal raising to increase IOP, extreme IOP in HBB raises questions about the nature of the gestures involved to produce ejectives and implosives in HBB. What are the gestures involved in raising and lowering the intraoral pressure for ejectives and implosives? In this study, we investigate the production of such sounds by means of MRI recordings of a single professional beatboxer. We think that (1) tongue root retraction is used to decrease the supralaryngeal volume for beatboxing ejectives and (2) tongue root advancement is used to increase the supralaryngeal volume for beatboxing implosives.

2. Methods

The subject (VP) of the study is a professional beatboxer. He was 32 years old at the time of the recordings and had practiced beatboxing for 15 years at the time.

2.1. Corpus & protocol

The corpus is composed of musical structures called "Beat Patterns" (BPs). They have same metric, rhythmic and melodic structure (see Figure 1). The metrics consists of 4 pulses and 9 sounds (musical notes). Rhythm is an alternation low/high, loud/soft and short/long sounds. Finally, the melodic parameters refer to the different timbres of the instruments. The phonetic structure of drums was manipulated to create 11 BPs in total. In this study, only 2 selected patterns are reported : [ɓ͡ʔ t͡sʼ ʔ͡ʃː t͡sʼ ɓ͡ʔ ɓ͡ʔ t͡sʼ ʔ͡ʃː t͡sʼ] and [ɗ͡ʔ t͡sʼ ↓ʔ͡ʃː t͡sʼ ɗ͡ʔ ɗ͡ʔ t͡sʼ ↓ʔ͡ʃː t͡sʼ] where [ɓ͡ʔ] and [ɗ͡ʔ] are (orally) unreleased implosives, respectively bilabial and dental; [t͡sʼ] is a dental ejective affricate and [ʔ͡ʃː] and [↓ʔ͡ʃː] are respectively pulmonic egressive and pulmonic ingressive affricates produced with an aryepiglottal stop and a post-alveolar fricative. The experiment consisted in repeating the audio examples of the *Beat Patterns* (BPs) of the corpus.

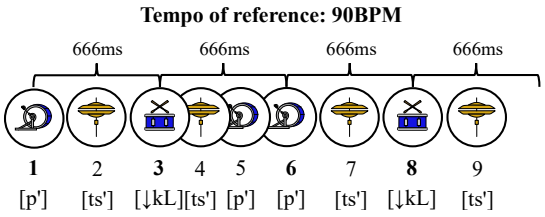


Figure 1: Beat pattern structure

2.2. Recordings

The MRI recording took place at Nancy Central Regional University Hospital (ClinicalTrials.gov Identifier: NCT02887053, RCB-ID n°CPP EST-III, 08.10.01). The data was acquired with a Siemens Prisma 3T scanner, Erlangen, Germany. The speaker was in supine position and a Siemens Head/Neck 64 coils was used. 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, in-plane resolution was 1.6 mm, recorded at 50 fps and reconstructed with a non-linear inverse technique presented in Uecker et al. (2010). 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) and the recording software presented in Isaieva et al. (2022).

2.3. Analysis

The MRI data was quantified with a functional Principal Component Analysis (fPCA) from the open source script of Belyk and McGettigan (2022) based on the semi-automatic MRI contouring procedure presented in Belyk, Carignan, and McGettigan (2023).

The onset and offset of articulatory gestures of the *throat kicks* [ɓ͡ʔ] and [ɗ͡ʔ] as well as the *closed hi-hat* [t͡sʼ] were selected for the analysis. The snares [ʔ͡ʃː] and [↓ʔ͡ʃː] were analyzed but are not reported here. Onset were defined as the moment of maximal constriction when the oral closure was achieved and articulators were not in motion anymore. Offset were defined as the end of the ejective or implosive phase, that is the frame when articulators were not in motion anymore to either compress or expand the pharyngeal cavity.

The semi-automatic contouring procedure was performed and manual correction of erroneous contours was applied. The procedure consists of mainly 4 steps :

1. frames were registered and head movement correction was performed based on a reference frame taken at the middle of the recording;
2. a mask was automatically created based on tissue detection in the reference frame capturing a region of interest from the lower edge of the larynx to the lips. The mask was manually corrected;
3. automatic contouring was performed based on the mask using pixel intensity to distinguish vocal tract edges (high intensity) from the air contained in the vocal tract (low intensity);
4. finally, contours were manually corrected when needed.

The resulting vocal tract contours (n= 128) represent the antero-posterior coordinates (Y-axis) and the supero-inferior coordinates (Z-axis). Contours were smoothed and centered at the lips with the *fda package* implemented in the script of Belyk and McGettigan (2022). Then, the fPCA analysis was performed on the Y and Z coordinates to find variation of contours in comparison to the mean shape of the vocal tract. The five first components explain 87% of the variation in the data :

- PC1 represents 37% of the variation and was found to be related to laryngeal height.
- PC2 represents 23% of the variation and was not found to be related to a specific articulatory gesture. It possible that opposite gestures "cancel" each other resulting in PC score similar to the mean shape.
- PC3 represents 15% of the variation and was found to be related to advancement or retraction of the tongue root.
- PC4 represents 6% of the variation and was found to be related to velopharyngeal compression or expansion.

- PC5 represents 6% of the variation and was found to be related to labial aperture and tongue blade constrictions. It mainly relates to the oral gestures to produce [ʔ͡ʃː] and [↓ʔ͡ʃː] and will not be reported here.

We plotted the vocal tract pattern of variation along the Y and Z coordinates associated to each principal component score. Boxplot representing the variation of the scores were plotted on R for each sound to interpret the graphical representation of vocal tract deformation.

3. Results

Figure 2 shows the results of the fPCA analysis for PC1 (laryngeal height), PC3 (tongue root maneuvers) and PC4 (velo-pharyngeal maneuvers). For each component, a plot of the vocal tract deformation relative to the mean shape is displayed. The mean shape is represented in black and the deformation is represented by contours in pink and green. Pink contours show the vocal tract deformation when the component increases. Green contours show the vocal tract deformation when the component decreases. Boxplots illustrate the components variation for each sound. Onset and offsets are represented in different colors and each panel illustrates BP. Although PC2 explains 23% of the variation in the data, it is not shown because in the graphic representing tract deformation for PC2, no deformation was observed and only the mean shape was visible.

The onset and offset of [ɓ̥] show differences in the vocal tract configuration. PC1, related to laryngeal height (Figure 2a), tends to be higher and positive at the onset. It suggests that larynx is in higher position at the onset. Though, laryngeal height seems somehow variable at the onset and some overlap is observed with the offset values on the boxplot. PC3, related to tongue root maneuvers (Figure 2b), increases from the onset to the offset suggesting systematic tongue root advancement. Finally, there is a participation of the velo-pharynx as suggested by PC4 (Figure 2c). The 4th component is higher at the offset and suggests the pharyngeal cavity is wide and the velum is raised. In the recordings, the velopharyngeal port remains closed.

The dental implosive [ɗ̥] is also characterized by tongue root advancement and pharyngeal expansion. PC3 is higher at the offset compared to the onset which confirms tongue root advancement (Figure 2b). The magnitude of tongue retraction seems weaker but less variable for the dental implosive compared to the bilabial one. PC4 is closed to 0 at the onset and increases at the offset suggesting the pharynx expands and the velum raises. Once again, the velopharyngeal port remains closed throughout the mechanism. PC1 suggests the larynx is lower at the onset and higher at the offset. Moreover, the difference between larynx height at the onset and offset is low. This suggests that laryngeal height is not that different between the onset and the offset. This result is not in line with the traditional view that implosives are produced by laryngeal lowering.

Concerning the *hi-hat* [t͡sʼ], differences are observed depending on the pattern, though tongue root retraction is systematic in both pattern. Indeed, PC3 is decreasing between the onset and the offset (Figure 2b) which suggests that tongue root is retracting. In BP4 (i.e. [ɓ̥ t͡sʼ ʔ͡ʃː t͡sʼ ɓ̥ ɓ̥ t͡sʼ ʔ͡ʃː t͡sʼ]) the difference between the onset and the offset is greater than in BP5 (i.e. [ɗ̥ t͡sʼ ↓ʔ͡ʃː t͡sʼ ɗ̥ ɗ̥ t͡sʼ ↓ʔ͡ʃː t͡sʼ]). The gesture has a greater amplitude in BP4, meaning the tongue root retracts more. In BP4, PC1 indicates that the larynx is higher at the offset. In BP5 the larynx is lower at the offset which is not in line with the traditional description of laryngeal raising during ejectives. Moreover in this BP, differences between laryngeal height at the onset and offset is very low suggesting laryngeal height is not that different. Finally, velo-pharyngeal involvement is observed in the 2nd BP but not in the first one. Indeed, the data on PC4 shows no major differences between the onset and the offset in BP4 while in BP5, the decrease of PC4 suggests velum is lowered and pharyngeal cavity is narrowed. Velum lowering does not mean the velo-pharyngeal port is opened, on the recordings it stays closed.

4. Discussion

The MRI analysis confirmed tongue root maneuvers during ejective and implosive mechanisms. Tongue root retraction was observed during ejectives, such action would result in increased intraoral pressure. Tongue root advancement was observed during implosives, such action would result in a negative pressure. Moreover, pharyngeal involvement was observed, suggesting a possible role of the pharyngeal muscles to increase or decrease the volume of the pharynx. However, the data does not suggest that laryngeal height is responsible for volumetric variation of the pharyngeal cavity.

Tongue root advancement and retraction are caused by different muscles activation. Tongue root advancement or protrusion is produced by activating the genioglossus (GG) (Shall 2012; Sanders and Mu 2013; Stone et al. 2018). Contracting the GG pulls the tongue toward the mandible.

Tongue retraction or retrusion is thought to be caused by the activation of the styloglossus (SG) and the hyoglossus (HG) *(ibid)*. Contraction of SG pulls the tongue upward and backward. HG activation pulls the tongue downward and backward. By activating both the SG and HG, the tongue root retracts.

A study of Saigusa et al. (2004) suggests that tongue root retraction is caused by the pharyngeal superior constrictor (SPC). They identified that some of the SPC fibers insert in the Transverse (T) muscle at the base of the tongue. They think that the activation of both the SPC and T fibers may result in large retraction of the tongue root and to pharyngeal constriction.

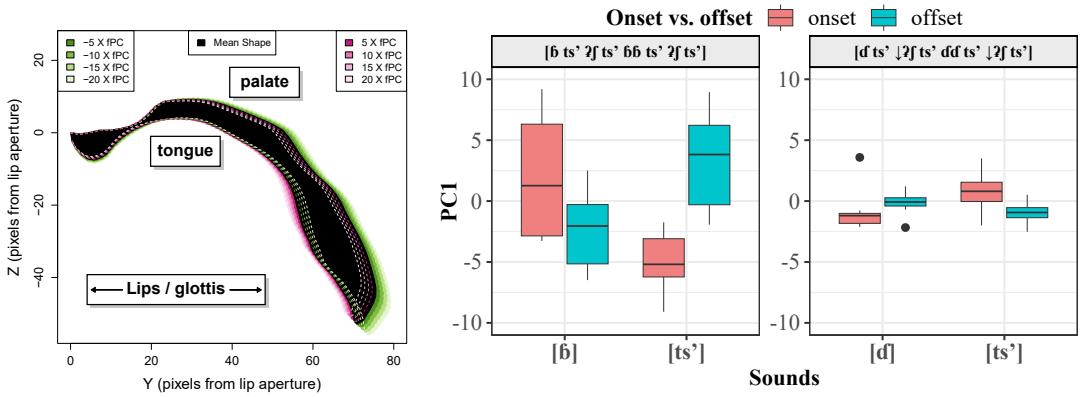
Tsumori et al. (2007) found that glossopharyngeal fibers of the SPC insert in the tongue root along with palatoglossus and styloglossus fibers. The authors suggest that the contraction of the SPC may play a role during swallowing

From a physiological point of view, Kokawa et al. (2006) investigated the production of the cardinal vowels /i a u/ based on naso-fibroscopy, X-ray fluorography and electromyography (EMG) of the SPC and GG muscles. Their findings suggest the activation of the SPC muscle retracts the root of the tongue while the activation of the GG protrudes the tongue root.

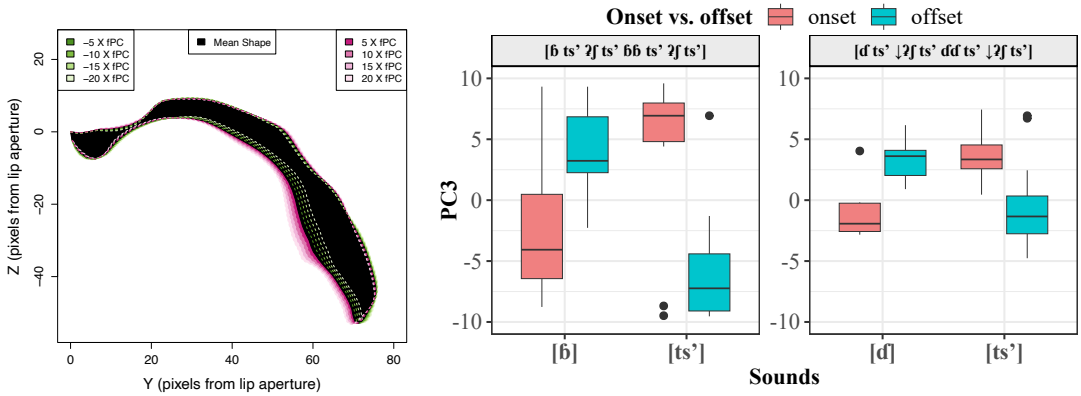
The action of the SPC muscle (and possibly the middle and inferior constrictors) may explain both tongue root maneuvers and differences in pharyngeal volume. If we consider that pharyngeal constrictors are responsible for volumetric variations, it is possible that laryngeal height is a mere consequence of contracting or relaxing pharyngeal constrictors.

5. Conclusion

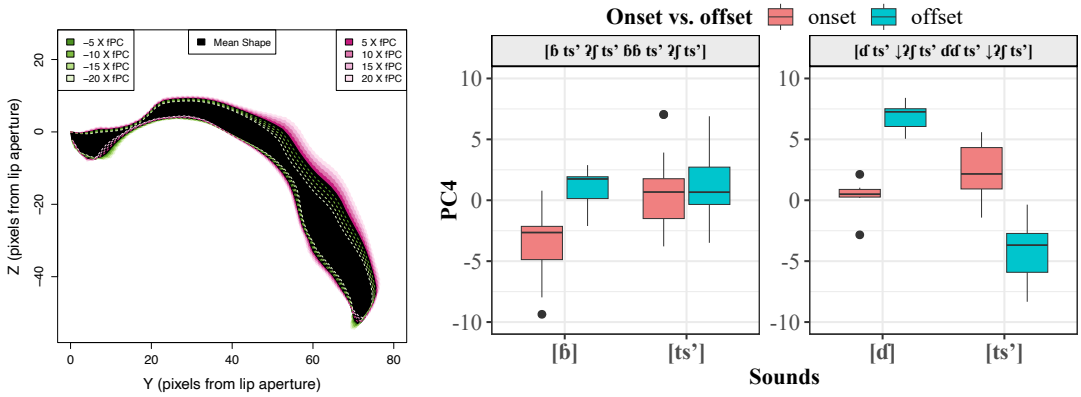
This study aimed to investigate the production of beatboxing ejectives and implosives. MRI recordings of a single professional beatboxer were analyzed via a fPCA analysis of vocal tract contours. The findings show that glottalic mechanisms, for this subject, are produced with tongue root and velo-pharyngeal maneuvers to change the volume of the pharyngeal cavity. Our results are not always in line with the textbook description of ejectives and implosives.



(a) Tract variation (left) and corresponding boxplot (right) of PC1 (laryngeal height).



(b) Tract variation (left) and corresponding boxplot (right) of PC3 (tongue root maneuvers).



(c) Tract variation (left) and corresponding boxplot (right) of PC4 (velo-pharyngeal maneuvers).

Figure 2: Left : Tract variation for fPC1, 3 and 4, black contours illustrate the mean shape of the vocal tract, pink contours indicate tract changes when a fPC increases while green contours indicates changes when a fPC decreases. Right : boxplot (n=96 frames) illustrating PC changes for each sound, colors indicate fPC variation between the onset (occlusion) and the offset (release).

Rather, we think ejectives and implosives might better be described as obstruent consonants produced with pharyngeal compression or expansion and a closed glottis (at least in beatboxing). A similar proposal has been made by Lindau (1979) who suggested to use the feature [expanded] pharynx for ATR vowels. She also proposes the features [neutral] [constricted] and [pharyngealized] to refer to differences in pharyngeal volume.

This novel working hypothesis needs to be verified on more beatboxers. It should also be tested on ejectives and implosives of the world's languages. If the hypothesis was to be confirmed, it would have several implications for the classification of stop and sound change.

6. Acknowledgements

This research was funded, in whole or in part, by LABEX EFL (Empirical Fundation of Linguistics, ANR-10-LABX-0083), the ArtSpeech project (ANR-15-CE23-0024) and supported by Eu-

ropean founds CPER “IT2MP”, “LCHN” and “FEDER”.

7. 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.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Bloomington: Indiana University Press. 278 pp.

De Torcy, Tiphaine, Agnès Clouet, Claire Pillot-Loiseau, Jacqueline Vaissière, Daniel Brasnu, and Lise Crevier-Buchman (Apr. 2014). “A video-fiberscopic study of laryngopharyngeal behaviour in the *human beatbox*”. In: *Logopedics Phoniatrics Vocology* 39.1, pp. 38–48. DOI: 10.3109/14015439.2013.784801.

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.

Dehais-Underdown, Alexis, Paul Vignes, Lise Crevier-Buchman, and Didier Demolin (2021). “In and out: production mechanisms in Human Beatboxing”. In: *Proceedings of Meetings on Acoustics* Vol. 45, No. 1. DOI: 10.1121/2.0001543.

Demolin, Didier (1995). “The phonetics and phonology of glottalized consonants in Lendu”. In: *Phonology and Phonetic Evidence*. Ed. by Bruce Connell and Amalia Arvaniti. Cambridge University Press, pp. 368–385. DOI: 10.1017/CBO9780511554315.026.

Demolin, Didier, Hubert Ngonga-Ke-Mbembe, and Alain Soquet (2002). “Phonetic characteristics of an unexploded palatal implosive in Hendo”. In: *Journal of the International Phonetic Association* 32.1, pp. 1–15. DOI: 10.1017/S0025100302000117.

Fabre, Christol (2018). “Les sons percussifs : des consonnes plosives au Human Beatbox, corrélations acoustiques, aérodynamiques et endoscopiques”. Master’s Thesis. Université Grenoble Alpes.

Hermes, Zainab, Maojing Fu, Sharon Rose, Ryan Shosted, and B. Sutton (2016). “Representations of Place and Airstream Mechanism : A real-time MRI study of Tigrinya ejectives”. In: LabPhon15. Ithaca, New York.

Isaieva, Karyna, Marc Fauvel, Nicolas Weber, Pierre-André Vuissoz, Jacques Felblinger, Julien Oster, and Freddy Odille (2022). “A hardware and software system for MRI applications requiring external device data”. In: *Magnetic Resonance in Medicine* 88.3, pp. 1406–1418. DOI: 10.1002/mrm.29280.

Kingston, John (1985). “The Phonetics and Phonology of the Timing of Oral and Glottal Events”. Doctoral Dissertation. University Of California. Berkeley.

Kokawa, Takayuki, Hideto Saigusa, Iichirou Aino, Chiharu Matsuoka, Tsuyoshi Nakamura, Kumiko Tanuma, Kazuo Yamashita, and Seiji Niimi (Sept. 2006). “Physiological Studies of Retrusive Movements of the Human Tongue”. In: *Journal of Voice* 20.3, pp. 414–422. DOI: 10.1016/j.jvoice.2005.08.004.

Ladefoged, Peter (1971). *Preliminaries to linguistic phonetics*. Midway reprints. Chicago, Ill.: University of Chicago Press. 122 pp.

Lindau, Mona (Apr. 1979). “The feature expanded”. In: *Journal of Phonetics* 7.2, pp. 163–176. DOI: 10.1016/S0095-4470(19)31047-2.

Oh, Miran and Yoonjeong Lee (Oct. 1, 2018). “ACT: An Automatic Centroid Tracking tool for analyzing vocal tract actions in real-time magnetic resonance imaging speech production data”. In: *The Journal of the Acoustical Society of America* 144.4, EL290–EL296. DOI: 10.1121/1.5057367.

Patil, Nimisha, Timothy Greer, Reed Blaylock, and Shrikanth S. Narayanan (Aug. 20, 2017). “Comparison of Basic Beatboxing Articulations Between Expert and Novice Artists Using Real-Time Magnetic Resonance Imaging”. In: *Interspeech 2017*. Interspeech 2017. ISCA, pp. 2277–2281. DOI: 10.21437/Interspeech.2017-1190.

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.

Saigusa, Hideto, Kazuo Yamashita, Kumiko Tanuma, Makoto Saigusa, and Seiji Niimi (2004). “Morphological studies for retrusive movement of the human adult tongue”. In: *Clinical Anatomy* 17.2, pp. 93–98. DOI: 10.1002/ca.10156.

Sanders, Ira and Liancai Mu (July 2013). “A Three-Dimensional Atlas of Human Tongue Muscles”. In: *The Anatomical Record* 296.7, pp. 1102–1114. DOI: 10.1002/ar.22711.

Shall, Mary Snyder (2012). “Tongue Biomechanics and Motor Control”. In: *Craniofacial Muscles*. Ed. by Linda K. McLoon and Francisco Andrade. New York, NY: Springer New York, pp. 229–240.

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. DOI: 10.1080/21681163.2016.1162752.

Sulaberidze, Nato, Erika Brandt, Phil Hoole, Martin Krämer, Jürgen R. Reichenbach, and Adrian Simpson (2023). “Ejectives in georgian. A real-time mri analysis of Vertical larynx movement”. In: *Proceedings of the 20th International Congress of Phonetic Sciences*, pp. 952–956.

Tsumori, Nobuaki, Shinichi Abe, Hiroko Agematsu, Masatsugu Hashimoto, and Yoshinobu Ide (2007). “Morphologic Characteristics of the Superior Pharyngeal Constrictor Muscle in Relation to the Function During Swallowing”. In: *Dysphagia* 22.2, pp. 122–129. DOI: 10.1007/s00455-006-9063-2.

Uecker, Martin, Shuo Zhang, Dirk Voit, Alexander Karaus, Klaus-Dietmar Merboldt, and Jens Frahm (2010). “Real-time MRI at a resolution of 20 ms”. In: *NMR in Biomedicine* 23.8, pp. 986–994. DOI: 10.1002/nbm.1585.

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

Abstract

Convergence, i.e. the process of two people adjusting their behaviour to become more similar to each other, has been found in most aspects of human interaction by now. However, most of the studies investigating convergence have so far considered mostly linguistic aspects while taking into account only production or perception rather than both. In this paper, we expand on previous work to examine paralinguistic phenomena, namely laughter, by integrating an analysis of differences between several acoustic cues extracted from laughter in spontaneous interaction with a perceptual experiment in order to determine their perceptual relevance.

Keywords: laughter, convergence, speech production, speech perception, conversational speech

1. Introduction

In recent years, many studies have investigated the behaviour of speakers in interaction and have found evidence of them adapting their communicative habits to become more similar to each other in a process known as convergence (Pardo 2013). These studies have shown convergence for syntactical (Branigan, Pickering, and Cleland 2000) and lexical (Brennan and Clark 1996; Nenkova, Gravano, and Hirschberg 2008) aspects of speech, among others. Segmental and prosodic features have also been examined for convergence (Pardo 2006; Pardo, Urmanche, et al. 2017; Levitan and Hirschberg 2011).

Some studies have also investigated paralinguistic items, such as pauses (Edlund, Heldner, and Hirschberg 2009), gestures (Holler and Wilkin 2011) and for conversational phenomena, such as laughter (Truong and Trouvain 2012a; Truong and Trouvain 2012b; Ludusan and P. Wagner 2019; Ludusan and P. Wagner 2022). Ludusan, Schröer, and P. Wagner (2022) previously showed, that speakers exhibit convergence of laughter regarding their vowel quality, as measured by comparing distances of F1 and F2 in between speakers at the start and end of the conversation.

Most phonetic studies of convergence can be grouped into two sets: One set studies convergence by calculating and comparing the distances in a given set of acoustic cues, such as spectral moments, formant values, etc. (Levitan and Hirschberg 2011; Pardo, Urmanche, et al. 2017; Gessinger et al. 2017; Ludusan, Schröer, and P. Wagner 2022) The other set looks at the phenomenon from a perceptual perspective by having listeners rate the similarity between audio stimuli (Pardo 2006; Pardo 2013; Babel 2012; Namy, Nygaard, and Sauerteig 2002).

Both of the aforementioned approaches provide important insights into the process of convergence, by either pinpointing acoustical cues that undergo statistical changes throughout the interaction or by showing that listeners perceive convergence to several different phonetic aspects.

Putting both approaches together should thus provide a more complete account of convergence by determining both the amount of change in the investigates acoustic measures, as well as their perceptual relevance. This holistic approach has started to become more popular in recent studies such as Abel and Babel (2017), M. Wagner et al. (2021), Lewandowski and Nygaard (2018), and Pardo, Jordan, et al. (2013).

In their study, Abel and Babel (2017) had pairs of participants perform a cooperative task. Afterwards, a different group of participants had to listen to and rate whether the pair had converged or not. While they found evidence for convergence both in the acoustic features extracted from the dyad itself as well as in the perceptual ratings, they could not establish a clear correlation between the two. They argued that, while the listeners had access to global changes across all acoustic dimensions, the acoustic difference algorithms they employed to analyse the data had access to only one dimension at a time. This would then suggest that not one cue might be important, but rather that perceptual information about convergence comes from an interplay between multiple cues.

Both M. Wagner et al. (2021) and Lewandowski and Nygaard (2018) integrated this in their studies investigating convergence towards both native and non-native accents. Each of these studies had participants produce a set of items and then shadow a different speaker going through the same list. The resulting productions were compared and the distance between several acoustic cues was calculated. They further presented both speakers' productions to a set of listeners who had to judge whether convergence took place (similar to Abel and Babel (2017)). Wagner et al. found vowel duration, speech rate and f0 to be the most converged-to and perceptually salient dimensions. The findings of Lewandowski and Nygaard further corroborate this, with one exception. In their study, the f0 measure only correlated with perceived convergence for non-native speakers converging to native speakers, while vowel quality was correlated with native to non-native convergence.

Pardo, Jordan, et al. (2013) had the same general setup as the previous studies, with a shadowing as well as an AXB perceptual task. While they also investigated lexical factors alongside acoustic ones, they could not establish a link between the lexical and perceived convergence. For the acoustic measures, however, they found vowel spectra, duration and f0 to correlate with perceived convergence, thus being in line with the findings of the two aforementioned studies

Here, we extend these findings by integrating production and perception aspects of laughter convergence in spontaneous interaction by testing the perceptual significance of several acoustic cues shown to differ within conversation.

2. Methods

2.1. Stimuli

Our stimuli were taken from the DUEL corpus (Hough et al. 2016), consisting of spontaneous dyadic interactions, in which speakers have to either role-play as a border control agent, discuss furnishing an apartment or write an embarrassing film script. In total, the German portion of the corpus contains 19 such dyads. One of the dyads was shown to have both speakers converge in Ludusan, Schröer, and P. Wagner (2022). As they produced far more instances of laughter in the film script condition than in the others, laughter from that condition was considered.

We segmented the laughter instances into syllables and vocalic/consonantal parts in accordance with Trouvain (2003) using VocalToolKit (Corretge 2022), a Praat (Boersma 2002) plugin. and corrected the annotations manually. Afterwards, we extracted vowels from the first and last third of the interaction, and we selected vowels which were egressive, artifact free, and non-fricated to be used as stimuli (n = 38). All but one of the extracted vowels were more than 100ms in length (on average 184ms). All steps were performed using Praat (Boersma 2002).

2.2. The perception experiment

Using the multiple forced choice experiment function provided by Praat (Boersma 2002), 23 participants were presented the stimuli in an AXB paradigm (Goldinger 1998; Pardo, Jordan, et al. 2013), modified so that the X was presented after A as well as after B, essentially resulting in two pairs (AX and BX). The participants (all German-speaking university students) were then asked to rate which pair was more similar. They were able to listen to each stimulus twice and could not go back in order to change their choice. At the beginning of each trial, there was a pause of 0.5s, so the trial would not start immediately after the previous answer was given. Furthermore, there was a pause of 300ms within and of 600ms between pairs in order to separate the stimuli and pairs clearly. The instructions on what the participants should listen for were left intentionally vague (only similarity was stated) in order to get as unbiased of a response as possible.

The stimuli had A taken from either the first or last third from one speaker, B taken from the last 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.

2.3. Analysis

We originally extracted the mean F1 and F2 values for every vowel used in the experiment in order to compute the Euclidean distance as described by Equation 1 and determine whether there was a change in vowel quality from the first to the last third for each speaker.

$$d_{f1f2}(A,B) = \sqrt{(F1_A - F1_B)^2 + (F2_A - F2_B)^2} \quad (1)$$

The formant values were normalised using the PhonR package (McCloy 2016) in R (R Core Team 2020), in which all other analyses were also carried out. We also examined, whether the convergence shown in Ludusan, Schröer, and P. Wagner (2022), for the dyad from which the vowels were extracted, could still be found in the stimuli used in this experiment by using a Wilcoxon test to test the difference in vowel quality between on speakers' baseline production and the other speaker at the start and end of the conversation.

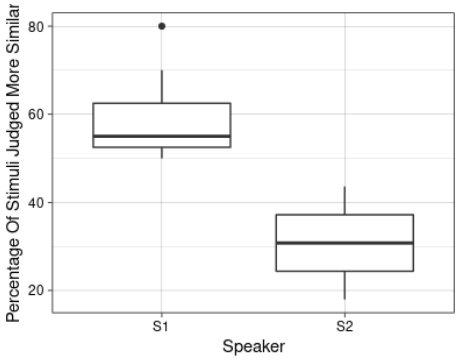


Figure 1: Percentage of how often each speaker in the last third of the interaction was perceived to be more similar to their interlocutors baseline than at the beginning of the conversation. A value significantly higher than 50% represents convergence, a value significantly lower than 50% shows divergence

However, since the preliminary results of the perceptual experiment and the acoustic analysis considering the F1F2 distances did not agree, we further extracted 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). For these features, we calculated absolute distances analogous to the example formula for f0 in Equation 2. We furthermore normalised all acoustic cues by subtracting their mean and dividing by two standard deviations (Gelman 2008).

$$d_{f0}(A,B) = |f0_A - f0_B|. \quad (2)$$

For the acoustic analysis Wilcoxon signed rank tests were used to test convergence, evaluating whether the distances between the speakers at the start of the conversation were greater than the distances between one speaker at the start (1st third) and one speaker at the end (last third) of the conversation. The distance between one speaker at the start and one at the end being smaller than the other indicates some degree of convergence of one speaker towards the baseline of the other, whereas the opposite would suggest divergence.

We fitted generalised mixed effects models in order to determine a link between perception and production, pooling the data from both speakers. The dependent variable was assigned a value of 1 if the raters had picked a pair consisting of vowels from different thirds as more similar (i.e. the convergence case) and was assigned a value of 0 in the divergence case.

All the examined cues were fixed factors in the model, and the rater was chosen as a random intercept. We employed model reduction by first building the largest possible model and then reducing it down step by step, as long as each step reduced the Akaike Information Criterion value of the model.

	dur		en		f0		cpp		F1F2	
	Speaker 1	Speaker 2	Speaker 1	Speaker 2	Speaker 1	Speaker 2	Speaker 1	Speaker 2	Speaker 1	Speaker 2
Mean	0.0126	-0.0071	-0.2739	1.6496	-7.1395	-88.5311	0.2015	1.6547	43.569	24.7816
SD	0.0226	0.0211	5.7367	4.0579	27.9043	65.096	2.5	2.3238	98.8131	97.84

Table 1: Means and standard deviation for the distance in difference value for each acoustic cue by which the speaker was being compared. Positive numbers indicate larger difference between the baselines, i.e. convergence.

3. Results

The Wilcoxon test performed prior to the perception study showed both speakers converging to their interlocutors' baseline over the course of the conversation in the F1F2 measure when considering all possible combinations of vowels in the stimulus set (S1 converging to S2 $p = 0.034$ and S2 to S1 $p = 0.0021$).

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%, $p = 2.8e^{-5}$)(s. fig. 1).

We then examined whether there are differences between 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 ($p = 0.008$), both indicating convergence. For speaker S2, the acoustic analysis showed a more complex picture, with the values for *en* ($p = 6.8e^{-4}$) and *cpp* ($p = 1.9e^{-4}$) showing convergence, while those for *dur* ($p = 0.033$) and *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 revealed 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.

4. Discussion

When taking into account instances between every combination of the data in the stimulus set, we found significant convergence for both speaker's vowel quality. This is in contrast to the results of the acoustical analysis performed only on the distances between the combination of token included in the perceptual experiment, in which the vowel quality measure was only significant for S1. This is due to the fact not every possible combination of stimuli was used, as this would have made the experiment exceedingly long. Thus, it is possible that vowel quality could have a more important role than shown by our study. 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 energy of the signal and the cepstral peak prominence measures indicated convergence, while f0 and duration 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 analysed 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.

While these results are based on a rather limited data set, they do align well with those of previous studies, that looked at both the production and perception aspects of convergence. As in Pardo, Jordan, et al. (2013) and M. Wagner et al. (2021) and Lewandowski and Nygaard (2018) duration, f0 and to a lesser extent the vowel quality seemed to play the most important role in determining whether listeners perceived convergence or not. Furthermore, it seems that as suggested in Pardo, Urmanche, et al. (2017) and Abel and Babel (2017) taking different cues and the interactions between them into account improves a model's ability to predict whether convergence/divergence are perceived. This further suggests that listeners not only have access to, but also integrate, several cues simultaneously in order to judge convergence. Our results further extend those of these studies, by showing that they hold for non-verbal phenomena as well.

In the future, we intend to address the limitations of this study by extending the analysis to a larger, more diverse dataset. It might further be interesting to try to incorporate even more acoustic cues. Lastly one may examine whether individual listeners vary in how they weigh acoustic cues, as well as whether different speakers tend to converge more strongly to different acoustic cues than others, as similar effects have been found for accent/speaker groups (Lewandowski and Nygaard 2018; M. Wagner et al. 2021)

5. Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 461442180.

6. References

Abel, Jennifer and Molly Babel (2017). "Cognitive Load Reduces Perceived Linguistic Convergence Between Dyads". In: *Language and Speech* 60.3, pp. 479–502. DOI: 10.1177/0023830916665652.

Babel, Molly (2012). "Evidence for phonetic and social selectivity in spontaneous phonetic imitation". In: *J. Phonetics* 40, pp. 177–189. DOI: https://doi.org/10.1016/j.wocn.2011.09.001.

Boersma, Paul (2002). "Praat, a system for doing phonetics by computer". In: *Glot International* 5, pp. 341–345.

Branigan, Holly, Martin Pickering, and Alexandra Cleland (2000). "Syntactic co-ordination in dialogue". In: *Cognition* 75, B13–25. DOI: 10.1016/S0010-0277(99)00081-5.

Brennan, Susan and Herbert Clark (1996). "Conceptual Pacts and Lexical Choice in Conversation". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, pp. 1482–1493. DOI: 10.1037/0278-7393.22.6.1482.

Corretge, Ramon (2022). *Praat Vocal Toolkit*. https://www.praatvocaltoolkit.com.

Edlund, Jens, Mattias Heldner, and Julia Hirschberg (2009). "Pause and gap length in face-to-face interaction". In: pp. 2779–2782. DOI: 10.21437/Interspeech.2009-710.

Gelman, Andrew (2008). "Scaling Regression Inputs by Dividing by Two Standard Deviations". In: *Statistics in medicine* 27, pp. 2865–73. DOI: 10.1002/sim.3107.

Gessinger, Iona, Eran Raveh, Sébastien Le Maguer, Bernd Möbius, and Ingmar Steiner (2017). "Shadowing Synthesized Speech-Segmental Analysis of Phonetic Convergence." In: *Interspeech*, pp. 3797–3801. DOI: 10.21437/Interspeech.2017-1433.

Goldinger, Stephen (1998). "Echoes of Echoes? An Episodic Theory of Lexical Access". In: *Psychological review* 105, pp. 251–79. DOI: 10.1037/0033-295X.105.2.251.

Holler, Judith and Katie Wilkin (2011). "Co-Speech Gesture Mimicry in the Process of Collaborative Referring During Face-to-Face Dialogue". In: *Journal of Nonverbal Behavior* 35.2, pp. 133–153. DOI: 10.1007/s10919-011-0105-6.

Hough, Julian, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg (2016). "DUEL: A multilingual 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 (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.

Ludusan, Bogdan and Petra Wagner (2019). "Laughter Dynamics in Dyadic Conversations". In: DOI: 10.21437/Interspeech.2019-1733.

— (2022). "Laughter entrainment in dyadic interactions: Temporal distribution and form". In: *Speech Communication* 136, pp. 42–52. DOI: doi.org/10.1016/j.specom.2021.11.001.

McCloy, Daniel (2016). *phonR: Tools for phoneticians and phonologists. R package version 1.0-7*. Online: https://cran.r-project.org/web/packages/phonR/phonR.pdf.

Namy, Laura L., Lynne Nygaard, and Denise Sauerteig (2002). "Gender Differences in Vocal Accommodation:: The Role of Perception". In: *Journal of Language and Social Psychology* 21.4, pp. 422–432. DOI: 10.1177/026192702237958.

Nenkova, Ani, Agustín Gravano, and Julia Hirschberg (2008). "High Frequency Word Entrainment in Spoken Dialogue." In: pp. 169–172. DOI: 10.3115/1557690.1557737.

Pardo, Jennifer (2006). "On phonetic convergence during conversation". In: *The Journal of the Acoustical Society of America* 119, pp. 2382–93. DOI: 10.1121/1.2178720.

— (2013). "Measuring phonetic convergence in speech production". In: *Frontiers in Psychology* 4, p. 559. DOI: 10.3389/fpsyg.2013.00559.

Pardo, Jennifer, Kelly Jordan, Rolliene Mallari, Caitlin Scanlon, and Eva Lewandowski (2013). "Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures". In: *Journal of Memory and Language* 69.3, pp. 183–195. DOI: https://doi.org/10.1016/j.jml.2013.06.002.

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.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Trouvain, Jürgen (2003). "Segmenting phonetic units in laughter". In: *Proc. of ICPhS*, pp. 2793–2796.

Truong, Khiet and Jürgen Trouvain (2012a). "Laughter Annotations in Conversational Speech Corpora – Possibilities and Limitations for Phonetic Analysis". In: *Proceedings of 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2012)*, pp. 20–24.

— (2012b). "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. DOI: 10.1016/j.wocn.2021.101076.

Phoneme monitoring and articulatory suppression in French-speaking adults

Claire Boilley[1,2], Patricia Pires[1], Hélène Lœvenbruck[2], 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, patricia.pires@etu-univ-grenoble-alpes.fr,
anne.vilain@univ-grenoble-alpes.fr, helene.loevenbruck@univ-grenoble-alpes.fr

Abstract

Segmenting words into their individual phonemes (explicit phonemic awareness) might require the use of abstract representations. To explore this hypothesis, we designed a phoneme monitoring task and examined how articulatory suppression and auditory interference affected performance in French-speaking adults with typical reading abilities. We controlled for the cognitive cost of dual-tasking by adding a semantic task. Contrary to our expectations, participants' scores were lower and reaction times longer under articulatory suppression, a finding that could not be attributed to interference from auditory feedback during articulation. We discuss potential linguistic and individual factors that might account for these unexpected results.

Keywords: phonological awareness, speech production

1. Introduction

Phonological awareness refers to the ability to detect and manipulate sublexical units in spoken words (Anthony & Francis, 2005). Performance in phonological awareness tasks, especially at the phoneme level (phonemic awareness), is an important predictor of learning to read (Melby-Lervåg, 2012). However, the ability to explicitly segment words into phonemes seems to emerge only after the onset of learning to read in an alphabetic system (Anthony & Francis, 2005). This has led some authors to claim that phoneme representations do not exist in pre-readers, or at all (Fowler et al., 2016). However, carefully designed tasks that require matching words based on a single shared phoneme instead of explicitly segmenting them show an early sensitivity to phonemes that does not appear to depend on letter knowledge (Ainsworth et al., 2019).

We suggest another interpretation: individuals who have not yet acquired any alphabetical knowledge struggle with tasks that demand explicit segmenting because they use a very concrete sensorimotor strategy, namely monitoring of their own covert articulation. Arguments for an articulatory strategy stem from the observation that pre-reading children perform better at segmenting a rime subsyllabic unit (VC) than a body (onset nucleus, CV). It has been suggested that segmenting a rime can be achieved by slowing down production and inserting a pause between the nucleus and the coda. However, inserting a pause within a prolonged CV sequence isolates the onset consonant, which is uncharacteristic of natural speech (Geudens et al., 2004). The articulatory features of individual phonemes also seem to interact with their position in predicting how difficult they are to segment (de Graaff et al., 2011). Because speech gestures for individual segments overlap within syllables, such a strategy may be more suitable for syllables than segments. In children, who generally coarticulate more than adults do, intra-syllabic coarticulation is inversely correlated with phoneme awareness (Noiray et al., 2019), which suggests that syllables are segmented earlier than phonemes by children because they are first produced as holistic undividable units.

On the other hand, evidence for abstraction in literate adults comes from a different type of task called internal phoneme monitoring, in which participants are instructed to judge the presence or absence of a given phoneme in words whose phonological form they covertly retrieve from viewing a picture, or using learned associations with other words. Using this task, Wheeldon & Levelt have shown that literate adults are hardly disturbed by concurrent articulatory suppression (Wheeldon & Levelt, 1995). Additionally, and even though reaction times to phonemes in different word positions showed that participants scanned words in a left-to-right fashion, these authors did not find any relationship between the overt articulatory length of a word's first syllable and latencies to detect the consonant at the onset of the second syllable. This mismatch reinforces the idea that subjects did not subvocalize the items during internal phoneme monitoring.

Here we replicate Wheeldon & Levelt's experiment with a sample of French-speaking adults, with the ultimate goal of adapting the task for children. We hypothesize that mastering the alphabetical principle allows one to use abstract phonological representations. Furthermore, we add an auditory interference condition to control for the effect of auditory feedback while articulating, and a semantic task to control for the effect of dual tasking. If literate adults indeed analyze abstract representations, we expect that articulatory suppression will have no significant effect once dual tasking and auditory feedback are accounted for.

2. Experimental study

2.1. Methods

2.1.1. Participants

The study was approved by the local ethics committee (Comité d'Éthique pour les Recherches de Grenoble Alpes, CERGA-Avis-2023-04). We tested 43 psychology students from Grenoble University (18 to 33 years old, F=37, M=5, other=1) who were native speakers of French, with normal or corrected-to-normal vision and hearing, and no history of speech or language disorder. Participants were rewarded with credits for their exams.

2.1.2. Task

The main experimental task was an internal phoneme monitoring task, in which participants were presented with an image, and were asked to tell whether a target phoneme was included in the word associated with the image, without overtly pronouncing the word. They provided their answer by pressing a button. This phonological task was performed in three different conditions: (1) the *simple* condition, where no concurrent task was performed, (2) the *articulatory suppression*

condition, in which subjects had to complete the task while continously repeating the non-word /bakusi/ aloud, starting before the presentation 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 conditions, a baseline semantic task was added, and performed in the same three conditions. The task consisted in looking at images and answering semantic questions about the images (e.g. "Is this a farm animal?") by pressing a button.

2.1.3. Stimuli

In the phonological task, the target phonemes 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. The carrier words were controlled for frequency and naming agreement (Duñabeitia et al., 2018). 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 complex CVC or CCV syllable (Cplx). Filler words for the phonological task and words for the semantic task were chosen from the same database as carrier words. We aimed to favor words comparable in length and structure to the carrier words, although this was not a strict criterion. Before the test, participants were familiarized with the pictures by viewing and naming (or if needed, repeating from the examiner) each of them once. Stimuli presentation and collection of responses and reaction times was conducted using the PsychoPy software (psychopy.org).

2.1.4. Procedure

The tasks were completed in a fixed order (semantic then phonological). Each task included three conditions, which were also presented in a fixed order (simple, articulatory suppression, and auditory interference). Starting with the simple condition allowed participants to become familiar with the task; auditory interference came last so subjects would not be influenced by any memory trace of the recorded interfering stimulus while performing articulatory suppression. In the phonological task, each condition comprised 2 blocks, with one target phoneme per block. The target phoneme was counterbalanced across 3 groups of participants so that each of the 6 phonemes was processed in each position and condition across the 3 groups. 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 keyboard. Each condition in each task started with a training block.

2.2. Data analysis and results

2.2.1. Data preparation

Data from three participants were excluded due to either not meeting inclusion criteria, or performing significantly lower than the rest of the group in the phonological task, suggesting potential phonological processing difficulties. Furthermore, we examined performance per item and found that the word "vulture", presented in the semantic task, clearly caused a higher number of errors than the other items; responses associated to this word were thus excluded from our data. For the analysis of reaction time, we used Wheeldon & Levelt's (1995) criterion and only retained responses that were correct and preceded by a correct answer. This was to ensure that response times were not affected by participants' awareness and processing of a previous error. Finally, we excluded responses with latencies more than 2 standard deviations above the individual participant means as calculated for each task and condition.

2.2.2. Data analysis and results

For each participant and condition, we calculated the difference between their performance on each item in the phonological task and their mean performance in the semantic (baseline) task. Analyses of accuracy and log-transformed reaction time to target words were then conducted with linear mixed-effects modeling (lme function in R), with condition as fixed effect and participant as random effect. Condition significantly predicted accuracy ($F=4.48$, $p=0.01$). Post-hoc multiple pairwise comparisons (emmeans function in R) revealed a significant difference between the simple and articulatory conditions ($t=2.79$, $p=0.02$) and a near-significant difference between the auditory and articulatory conditions ($t=-2.34$, $p=0.06$), but no significant difference between the simple and auditory conditions ($t=0.45$, $p=0.89$) (Figure 1).
Likewise, reaction time (Fig.2) was significantly affected by condition ($F=10.06$, $p<.001$). Pairwise comparisons showed, again, a significant difference between the simple and articulatory conditions ($t=-3.91$, $p<.001$), but also between the simple and auditory conditions ($t=-3.78$, $p<.001$), while the difference between the articulatory and auditory conditions was not significant ($t=0.25$, $p=0.97$).

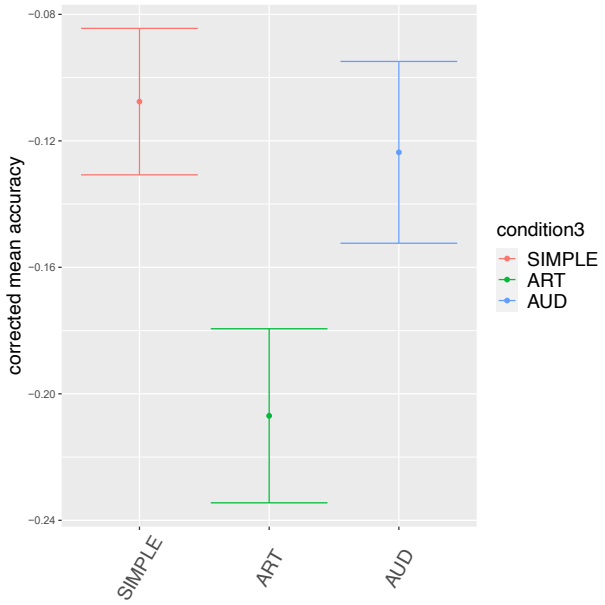


Figure 1: *Average difference in accuracy between phoneme monitoring and baseline semantic tasks across the different conditions. Note that the values are negative because on average, accuracy was higher in the semantic than in the phonological task.*

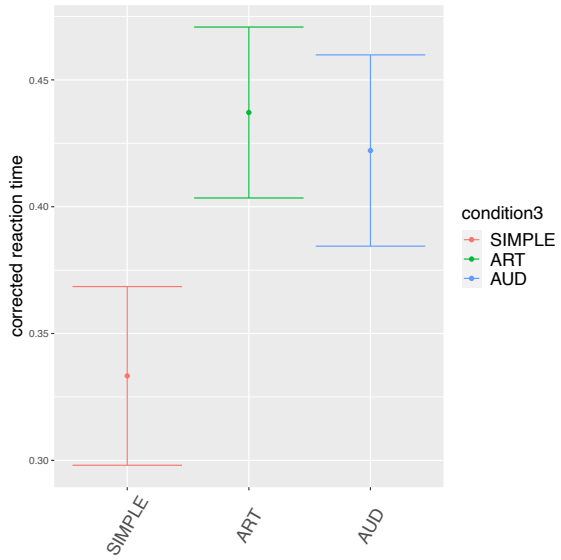


Figure 2: *average difference in reaction time between phoneme monitoring and baseline semantic tasks across the different conditions.*

There was no significant interaction between phoneme position (C1, C2, C3, Cplx) and condition with regard to accuracy or reaction time.

2.3. Discussion

We found that articulatory suppression significantly lengthens reaction time and lowers accuracy of internal phoneme monitoring in French-speaking adults, compared with a simple, unperturbed condition, even when controlling for the added cognitive cost of performing a dual task. This is contrary to previous results (Wheeldon & Levelt, 1995), and to the assumption that participants use a strategy based on abstract phonological representations. Furthermore, it could be argued that auditory feedback contributes to the effect of articulatory suppression on reaction time. However, this may not be the case for accuracy, as the auditory interference condition resulted in longer reaction times (similar to those observed with articulatory suppression) but did not lead to lower accuracy scores. This suggests that the effect of articulatory suppression on accuracy may not be solely due to auditory feedback, and that other factors may be at play. The effects of condition were consistent across all target positions.

The observed effects of articulatory suppression and auditory interference on phoneme monitoring in this covert naming (inner speech) task provide evidence that both articulatory and auditory representations may be involved in this task. Current models of inner speech have different views on the nature of representations involved in covert naming. While some researchers propose that inner speech is impoverished at the featural (articulatory and auditory) level (e.g. MacKay, 1992; Oppenheim & Dell, 2008), others suggest that inner speech involves both auditory and motor representations. One influential theoretical model proposes two streams for inner speech production (Tian & Poeppel, 2012). According to this model, auditory word forms could be either directly retrieved from memory, or reconstructed from the predictions of auditory consequences of simulated articulatory gestures. In this framework, "imagined hearing" is supposed to involve the memory retrieval route, whereas "imagined speaking" involves the motor simulation route .(Tian et al., 2016). Our findings of an involvement of articulatory and auditory representations are compatible with the motor simulation route. Another view suggests that inner speech varies between condensed forms, in an abstract phonological format, and expanded forms that include full somatosensory and auditory representations, derived from an efference copy mechanism with multisensory prediction (Grandchamp et al., 2019). Our present findings are also compatible with this view, suggesting that articulatory simulation and auditory representations may be involved in inner speech processing. Further, recent research suggests that the motor system may play a role in identifying phonetic elements but not in computing their sequence in words (Berent et al., 2023).

Why did Wheeldon & Levelt (1995) not find significant effects of articulatory suppression? Although methodological differences could be at play, such as the characteristics of the familiarization phase or the presence of a control condition, the potential role of orthographic representations should also be discussed. Although we focused our analysis on the comparison between the phonological and semantic tasks (with a difference score), the raw scores in the phonological task alone reveal relatively high error rates, averaging around 10% in the simple condition. This is considerably higher than in Wheeldon et al. (1995), or Manoiloff et al. (2015), who both obtained an error rate under 5%. These authors worked with speakers of Dutch and Spanish, respectively. Interestingly, to our knowledge, error rates comparable to ours were only observed in a phoneme monitoring study with adult English-speaking participants (Howell & Bernstein Ratner, 2018). French and English happen to have the most opaque orthographic systems of all major Western European languages, with a demonstrated impact on the development of phonological awareness and reading. Speakers' strategies might vary depending on the depth of their languages' orthographies. Although we tried to control for item orthographic complexity and found no significant effect, our study design made it difficult to investigate such effects without losing statistical power.

Another possibility lies in prosodic factors. Although both Dutch and English favor certain lexical stress patterns, French tends to stress the last element of a prosodic group that may be a word or a larger entity (Jun & Fougeron, 2000; Payne, 2021). The intonation contour of an isolated word may thus appear "flatter" in French than it is in English or Dutch. In speech perception, stress may help process the structure of elements within a syllable, possibly by supporting adequate temporal alignment between auditory cortical oscillations and the acoustic signal of speech (Goswami, 2022). If French-speaking individuals cannot readily rely on this clue, then (simulated) articulation may be necessary in order to sequence phonemes accurately, making phoneme monitoring less resilient to articulatory suppression.

Assuming our participants used articulatory representations to complete the phoneme monitoring task, one must then explain why auditory interference without concurrent articulation significantly slowed reaction times. It could be the case that hearing interfering speech activates the cortical network involved in speech production. Such motor activations support the identification of speech sounds in challenging listening conditions such as the presence of background noise (Wu et al., 2014). In our experiment, the interfering stimulus was a pseudo-word, repeated in an infinite loop with rather unnatural prosody (due to the absence of breaks), which may have constituted difficult enough listening conditions for participants to activate their speech motor network while accessing and analyzing inner representations. This concurrent activation might have slowed down the mental scanning of items. However, this did not affect response accuracy, suggesting either that our participants had

become familiar enough with the interfering non-word (which they had uttered many times under the preceding articulatory suppression condition), or that only the individual's own speech plans can lower the accuracy of phoneme monitoring.

Alternatively, this difference between articulatory suppression and auditory interference could reflect the use of temporal strategies. Anecdotally, experimenters witnessed high variability in participants' actual behavior during articulatory suppression. In particular, some systematically inserted short breaks between successive iterations of the pseudo-word, or slowed down while processing certain items, despite repeated instruction to keep a steady and smooth rate, as modelled by the experimenters. Slowing down or interrupting articulation might allow allocation of resources to the current item. Using this strategy, however, is not possible during auditory interference, because the interfering stimulus is played in a continuous loop, which the subject cannot control.

Finally, another interpretation, which is compatible with the models of Tian & Poeppel (2012) or Grandchamp et al. (2019), is that phoneme monitoring is ultimately based on multisensory representations (auditory and somatosensory) that are derived from motor simulation. In the auditory interference condition, the auditory predictions generated by motor simulation are masked by the interfering word, leading to slower responses. However, the accuracy of responses is not affected, as somatosensory elements may compensate for the degraded auditory representation. In the articulatory suppression condition, both the somatosensory and auditory representations generated by motor simulation are disrupted, leading to slower and less accurate responses. This interpretation highlights the importance of multisensory representations in inner speech processing, and suggests that motor simulation may play a critical role in generating these representations.

3. Conclusion

In conclusion, our data do not allow us to claim that French-speaking literate adults use abstract representations when monitoring inner word representations for specified phonemes. Linguistic factors such as orthographic transparency or prosody might explain some differences with previous experiments. Finally, behavioral diversity in articulatory suppression suggests the need to investigate the potential effects of individual differences in temporal strategies. Future work should aim to better understand these factors and to examine the effects of articulatory suppression in children at varying stages of reading acquisition.

4. Acknowledgements

We would like to thank Morgane Rossignol for her help with data collection and Silvain Gerber for his precious advice on statistical analysis. This project was supported by funding from the French National Agency for Research (ANR-19-CE28-0016)

5. References

Ainsworth, S., Welbourne, S., Woollams, A., & Hesketh, A. (2019). Contrasting Explicit With Implicit Measures of Children's Representations : The Case of Segmental Phonology. *Language Learning*, *69*(2), 323-365.

Anthony, J. L., & Francis, D. J. (2005). Development of Phonological Awareness. *Current Directions in Psychological Science*, *14*(5), 255-259.

Berent, I., Fried, P. J., Theodore, R. M., Manning, D., & Pascual-Leone, A. (2023). Phonetic categorization relies on motor simulation, but combinatorial phonological computations are abstract. *Scientific Reports*, *13*(1), 874.

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.

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic : A standardized set of 750 drawings with norms for six European languages. *Quarterly journal of experimental psychology*, *71*(4), 808-816.

Fowler, C. A., Shankweiler, D., & Studdert-Kennedy, M. (2016). « Perception of the speech code » revisited : Speech is alphabetic after all. Psychological Review, 123(2), 125-150.

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.

Goswami, U. (2022). Language acquisition and speech rhythm patterns : An auditory neuroscience perspective. *Royal Society Open Science*, *9*(7), 211855.

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, Dialogicality and Intentionality dimensions of inner speech within a hierarchical predictive control framework. Frontiers in psychology, Exploring the Nature, Content, and Frequency of Intrapersonal Communication, 10.

Howell, T. A., & Bernstein Ratner, N. (2018). Use of a phoneme monitoring task to examine lexical access in adults who do and do not stutter. *Journal of Fluency Disorders*, *57*, 65-73.

Jun, S. A., & Fougeron, C. (2000). A phonological model of French intonation. In *Intonation: Analysis, modelling and technology* (pp. 209-242). Dordrecht: Springer Netherlands.

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.

MacKay, D. G. (1992). Constraints on theories of inner speech. In D. Reisberg (ed.), *Auditory Imagery*. NJ/England: Erlbaum, 121–49.

Noiray, A., Popescu, A., Killmer, H., Rubertus, E., Krüger, S., & Hintermeier, L. (2019). Spoken Language Development and the Challenge of Skill Integration. *Frontiers Psychol.*, *10*, 2777.

Oppenheim, G. M., & Dell, G. S. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, 106, 528–37.

Payne, E. (2021). Comparing and deconstructing speech rhythm across Romance languages. In *Manual of romance phonetics and phonology* (p. 264-298). De Gruyter.

Tian, X., & Poeppel, D. (2012). Mental imagery of speech : Linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neurosci.*, *6*.

Tian, X., Zarate, J. M., & Poeppel, D. (2016). Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex*, *77*, 1-12.

Wheeldon, L. R., & Levelt, W. J. M. (1995). Monitoring the Time Course of Phonological Encoding. *Journal of Memory and Language*, *34*(3), 311-334.

Wu, Z.-M., Chen, M.-L., Wu, X.-H., & Li, L. (2014). Interaction between auditory and motor systems in speech perception. *Neuroscience Bulletin*, *30*(3), 490-496.

Temporal coordination of articulatory and respiratory events prior to speech initiation

Oksana Rasskazova[1], Christine Mooshammer[1], Susanne Fuchs[2]

[1]*Humboldt-Universität zu Berlin, Germany*

[2]*Leibniz-Centre General Linguistics, Germany*

oxanarass@gmail.com, christine.mooshammer@hu-berlin.de, fuchs@leibniz-zas.de

Abstract

Understanding the temporal organization of articulatory and respiratory events prior to speech initiation is crucial for encoding the complexities of speech planning and speech production. This study builds on a pilot study (Rasskazova et al. 2019), which demonstrated temporal alignment between oral articulators and exhalation onset for alveolar consonants of 6 speakers. In the current study, we investigate 11 speakers, five initial segments and two sentence conditions: utterance-initial silent interval and inter-speech pauses. Our results indicate a tight coupling between the onset of exhalation with the velocity peak of the closing gesture and the nucleus onset of the initial segment, shown by latencies with very low variability. This suggests a synchronized timing between respiratory and articulatory events. Preparatory events were more variable during utterance-initial positions compared to inter-speech pauses.

Keywords: oral-respiratory coordination, articulatory speech planning, EMA, respitrace

1. Introduction

During speech pauses or silent intervals prior to an utterance, speakers plan the upcoming speech at various levels (e.g., Krivokapić 2014). The complexity of how speech planning is orchestrated remains largely unclear. Planning on the phonetic level is characterized by preparing the motor actions, i.e., articulatory gestures for the initial speech segments (e.g., Rastle et al. 2005; Ramanarayanan et al. 2009; Mooshammer et al. 2012) as well as other speaker-specific preparatory activities such as speech-ready posture as highlighted in studies by Gick et al. (2004), Krivokapić et al. (2020), and Rasskazova et al. (2018). The results of these studies indicate that speech preparation involves temporally coordinated vocal tract actions in silent speech intervals (see also Rasskazova et al. 2019). The results of previous studies on temporal aspects of speech planning show that the movements of the articulators usually start well before the acoustic onset. In highly controlled naming tasks, this delay varies between 120 -180 ms and depends on the manner of articulation of the initial segment (Mooshammer et al. 2012; Schaeffler et al. 2014). In spontaneous dialogues, articulatory anticipation for very short turns can start as early as 3 seconds before the acoustic onset (Krause et al. 2021). Respiration may play a crucial role in the temporal organization of speech preparation. First, empirical results for this relationship come from studies on respiratory activities at the beginning of an utterance (Slifka 2003; Fuchs, Petrone, et al. 2013; Zöllner et al. 2021). Moreover, respiratory dynamics, particularly variations in inhalation duration and depth, might play a critical role in speech planning. This aspect of planning seems to be associated with the anticipated length of the forthcoming utterance rather than with the identity of initial segments (Whalen et al. 1997; Fuchs, Petrone, et al. 2013). Previous studies, such as Rochet-Capellan et al. (2014), have suggested a significant linkage between the acoustic onset of speech and the beginning of exhalation, yet a comprehensive understanding of the coordination with vocal tract gestures is still lacking. To our knowledge, respiration has not yet been included as an active gesture in Articulatory Phonology (cf. Browman et al. 1992) and Task dynamics (Saltzman et al. 1989), although exhalation is essential to speech (with the exception of very short utterances and paralinguistic ingressive speech). In our previous study (Rasskazova et al. 2019), we presented preliminary results on the coordination of respiratory, acoustic and articulatory events prior to the utterance for the alveolar consonants /t/ and /n/ for six speakers. We found evidence for temporal alignment between oral articulators and the onset of exhalation. The articulatory initiation of the initial segments started during the final phase of the inhalation, which was almost synchronous with the nucleus onset. This timing tended to be sensitive to the identity of the initial segment and speaks for a close coordination between respiratory and articulatory actions. The current study extends the investigation of Rasskazova et al. (2019) to 11 speakers, five initial segments /t/, /n/, /ʃ/, /a/, /h/ as well as two types of silent pre-speech intervals: the utterance initial (before the first utterance) and inter-speech interval (between the first and second utterance).

2. Methods

2.1. Participants and Material

Eleven native German speakers (5 male, 6 female), aged between 22 and 38 years, without a known history of respiratory or articulatory disorders and hearing impairment, participated in the study. The participants performed a reading task. The speech material involved eleven two-sentence combinations, which were repeated 5 times in randomized order. We included several filler sentences, which differ in structure and consist of one sentence only. The utterances were presented on a computer screen. The target utterance was controlled for the initial segment, sentence length and word stress. Each sentence was between 22 and 25 syllables and started with /a/, /t/, /n/, /h/ or /ʃ/. Therefore, we could compare the coordination of respiratory and articulatory events in utterance initial silent interval and inter-speech pause.

2.2. Recording procedure

Respiration, speech kinematics and acoustics were simultaneously recorded by means of Electromagnetic Articulography (EMA (AG501)) and Inductance Plethysmography. Acoustic data were recorded at 44.1 kHz using a shotgun microphone. The EMA sensors were attached to the tongue tip (TT), tongue middle (TM) and tongue back (TB), the jaw, and the upper and lower lips (UL, LL). Four reference sensors were included to compensate for head movements. The articulatory data were recorded at a sampling rate of 1250 Hz and then downsampled to 250 Hz post-processing in MATLAB. The data were corrected for head movement and then rotated and translated to the bite plane or to the fictional plane between the upper incisors and the nose. To record the respiration data two elasticized bands, one around the ribcage and another around the abdomen, were applied to the participants. The two respiratory signals were recorded simultaneously with the audio signal on a multi-channel DAC 6B recorder. The Inductance Plethysmography system was connected to the EMA system by means of a synchronization box. The synchronization impulse of the EMA was recorded on the multi-channel DAC recorder. Before each trial, participants heard a beep signal. The participants received this signal as a trigger so that they could start reading the presented speech material. The onset of the beep signal as well as the time-point of speech onset were recorded, but not controlled.

2.3. Measurements

The acoustic, articulatory and respiratory data are labelled with the visualization and labelling tool MVIEW (Tiede 2005), written in MATLAB. A custom-made labelling procedure detects acoustic onsets for reaction-time data following utterance based on the RMS peak amplitudes. Temporal respiratory events were labelled as inhalation minima for the onset of inhalation and maxima for the onset of exhalation. The sum of the two signals was taken for the labeling procedure. This labelling procedure was applied to the utterance initial and inter-speech silent intervals. The results of automatic labelling were carefully checked and, if needed, manually corrected. The movement onset of the lower lip and the articulatory gestural phases of the initial speech segment are determined automatically by using a 20% threshold criterion of the tangential velocity signal. To label the movement of articulatory phases, the tongue tip signal (TT) was labelled for the initial alveolar consonants /t/, /n/, /ʃ/ and the tongue back (TB) for the vocalic gestures of /a/ and /h/. To investigate temporal coordination, the following events prior to the first utterance as well prior the second utterance (inter-speech pause) are analyzed: acoustic onset of speech, inhalation and exhalation onset, movement onset of the lower lip for mouth opening as well the tongue tip gesture of the initial segment of the utterances (cf. Figure 1). The exhalation onset was used as a reference point and all events were subtracted from it. To test whether the latencies are affected by sentence position and initial segment linear mixed effect models were calculated (Bates et al. 2015). To test the stability of the latencies, we calculated the Relative Standard Deviation (RSD) using R 4.2.1 (R Core Team 2022), assuming that latencies with less variability show stable coordination between events.

3. Results

Based on results from Rasskazova et al. (2019) we expect that the onset of exhalation is coordinated with the nucleus onset of initial segments. Furthermore, and more exploratory, we assume that the manner of articulation of the initial segment as well as sentence position affects the timing between oral and respiratory events.

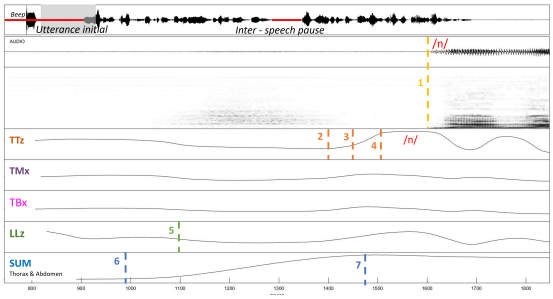


Figure 1: *Labelling procedure of measuring acoustic, articulatory and respiratory parameters for initial /n/ in utterance initial position. The acoustic signal was labelled for the acoustic onset (1). The tongue tip (TT) signal was labelled for gesture onset of /n/ (2), peak velocity (3) and nucleus onset. The lower lip gesture (LL) was labelled for its onset (7). The respiratory signal (SUM) was labelled for the inhalation (6) and exhalation (7) onsets.*

3.1. Temporal organization of acoustic, articulatory and respiratory events

Figure 2 shows the latencies of acoustic, respiratory and articulatory events, averaged across speakers and initial segments, during the utterance initial silent interval (red) and inter-speech pause (blue). The inhalation in both sentence conditions starts as the first preparatory event. The inhalation duration is on average 760 ms in utterance initial condition and 496 ms during the inter-speech pause. The movement onset of the lower lip starts 499 ms and 373 ms sentence position, respectively. The gestural onset for the initial segment always starts before the exhalation with an average of 116 ms for utterance-initial and 108 for inter-speech position, followed by peak velocity and nucleus onset. The acoustic onset starts shortly after the beginning of exhalation.

To investigate the effects of sentence position and manner of articulation of initial segments on the timing of preparatory events we calculated linear mixed effect models. The model was run for each latency with two fixed effects and subject as a random factor. Additionally, multiple comparison tests were run to identify which groups of articulatory types differ from each other. For both, the onset of inhalation and the movement onset of the lower lip, a significant effect ($p < .0001$) of sentence position was found, i.e. during utterance-initial position the speakers start to inhale and open the lips much earlier than during inter-speech pauses (cf. Figure 2), leading to shorter inhalation phases in inter-speech pauses. The latencies of the gestural onset showed significant differences for the initial segment between /ʃ/ and vocalic /h/ ($t=3.8$, $p<0.001$). Thus, /ʃ/ starts closest to the exhalation onset around 83 ms prior to the exhalation begin. The alveolar consonants /t/ and /n/ are almost identical in their timing 101 ms and 106 ms prior to the exhalation. The vocalic gestures /a/ and /h/ start much earlier, especially onset of /h/ starting 148 ms prior to exhalation. The sentence position does not affect the timing significantly as a main effect. However, the vocalic gestures start earlier in utterance-initial

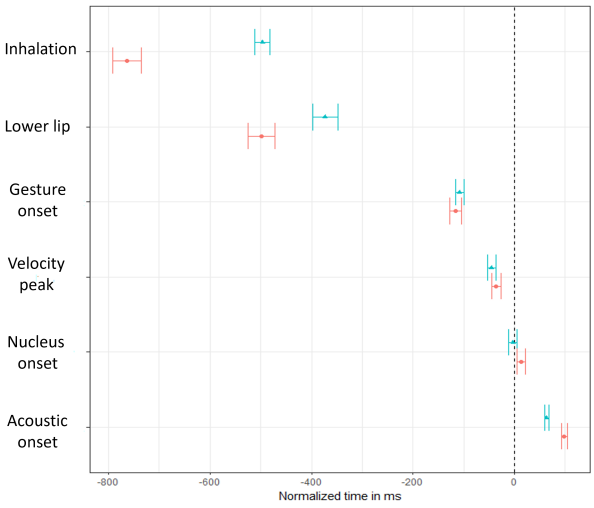


Figure 2: *Means and standard deviations of latencies relative to the exhalation onset (line-up point at 0 on x-axis) for the speech initiation during utterance-initial silent interval (red) and inter-speech pause (blue). Preparatory events are shown on the y-axis.*

position than during the inter-speech pause. The latency of the velocity peak of the constriction gesture for initial /ʃ/differs significantly from /n/ (t=3.6, p<0.002) and from vocalic /h/ (t=3.0, p<0.002). The average velocity peak latency for /ʃ/ in utterance initial position is exactly 0, i.e. synchronous with exhalation onset. For other segments, the peak velocity lies on average between 30 and 55 ms prior to the exhalation onset (cf. Figure 3).

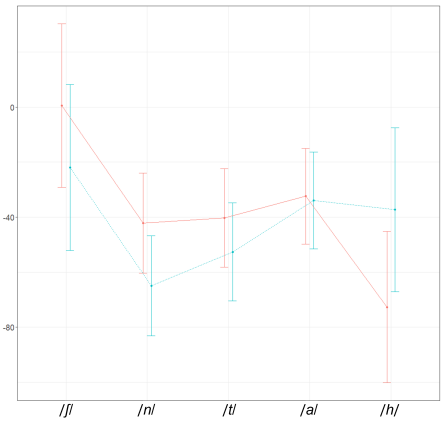


Figure 3: *Modelled results of peak velocity latencies during utterance-initial silent intervals (red) and inter-speech pauses (blue) for the initial segments (x-axis). The latency in ms is shown at the y-axis with 0 as exhalation onset.*

For initial segments /t/ and /n/ as well vocalic /h/ the nucleus onset starts almost synchronously with exhalation onset, with a delay of only 20 ms. The nucleus onset latency of the fricative /ʃ/ significantly longer than for /n/ and /h/ (t=6.2, p<.0001,), with a values of average 47 ms after exhalation onset. Another significant difference is observed between consonants /t/ and /n/ and the vowel /a/ (t=7.5,p<.0001), the difference in timing of the nucleus onset between those segments is on average 55 ms.

Overall, the timing for the nucleus onset of the alveolar consonants /t/ and /n/ is identical, being almost synchronous with exhalation onset. For /ʃ/ and vocalic gestures /a/ and /h/ the nucleus onset starts after the exhalation. The acoustic onset happens with a short average delay of 89 ms after the onset of exhalation in utterance-initial position and 64 ms during inter-speech pause. The fricative /ʃ/ differs significantly from the other initial segments (p<.0001) occurring closer to the exhalation onset: 52 ms in the utterance-initial position and only 24 ms during inter-speech condition. This difference is also significant with regard to sentence position (p<.0001) (cf. Figure 4).

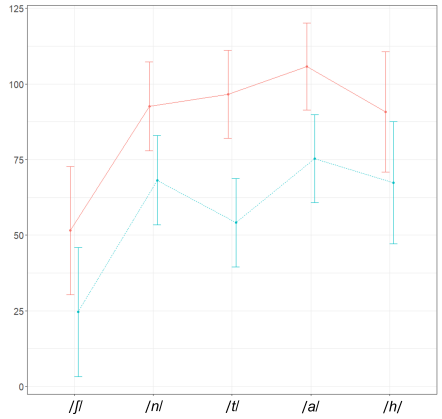


Figure 4: *Modelled results of acoustic onset latencies during utterance-initial silent interval (red) and inter-speech pause (blue) for the initial segments (x-axis). The latency in ms is shown on the y-axis with 0 as exhalation onset.*

3.2. Variation

Figure 5 shows that inhalation is by far the most variable preparatory event across all speakers. In utterance-initial position, its relative standard deviation (RSD in %) is 24% and 10% in inter-speech pauses. The onset of the lower lip movement is the second most variable latency with 17% and 13% variation in the respective sentence position. The phases of articulatory gestures seem to be stable across all speakers and initial segments. The gesture onset latency is the most variable (6.5%) in utterance-initial position, following peak velocity of the closing gesture (5%) and nucleus onset with 4.6 %. During the inter-speech condition, the variation of gestural phases is smaller and more consistent ranging from 4.6% for gestural onset and 4.2% for both peak velocity and nucleus onset. The initial segment /ʃ/ shows in both sentence conditions the least variability. The highest variability was found for the initial segment /h/ in utterance initial position and the least variability for the acoustic onset with only 3.3% in utterance-initial and 2.4% in inter-speech condition.

4. Discussion

In this study, we found that inhalation onset, relative to exhalation, was affected by sentence position: During utterance-initial position, the inhalation onset starts much earlier and is also more variable than during inter-speech pauses. This finding cannot be explained by the length of an upcoming utterance, since both sentences were controlled for their length and they varied between 22 and 25 syllables. Inter-speech pauses

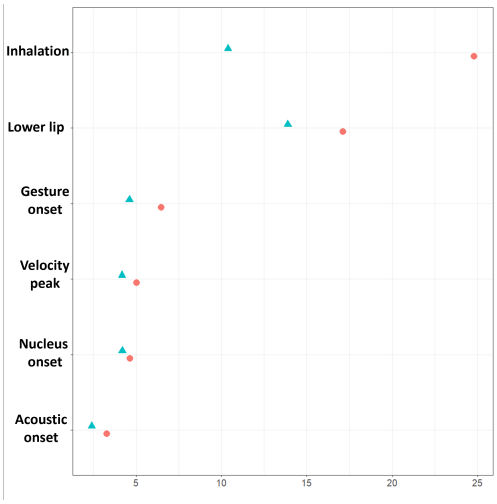


Figure 5: *Average Relative Standard Deviation in % for each latency (y-axis) during utterance-initial silent interval (red) and inter-speech pause (blue).*

are shorter and the speaker might compensate for the lack of time by the volume of inhalation, i.e. during utterance-initial pauses speakers have time and inhale rather slowly, but during inter-speech pauses, they inhale fast. To gain more insights into this effect, additional data on inhalation depth and velocity needs to be included in the analysis. Inhalation is also the most variable preparatory event. This indicates that speakers can adapt their physiological needs like inhalation to the speech context. Using an extended dataset, we can confirm our results from Rasskazova et al. (2019): the onset of expiration is tightly coupled with articulatory gestures as shown by the low variability of articulatory latencies. The coupling is sensitive to the identity of the initial segment, but not to sentence position. For all initial segments, speakers anticipate the first gesture before exhalation onset. The nucleus onset is almost synchronous with the exhalation onset for the alveolar consonants /n/ and /t/ while for the fricative /ʃ/ the closing peak velocity is exactly synchronous with exhalation onset. For vowels, both latencies are starting right after the exhalation. For the fricative, an early achievement of the nucleus might be essential for building up the necessary airflow to generate noise. For all segments and sentence conditions, the acoustic onset has the smallest variability. This finding is especially interesting since it might indicate that the planned target is not the articulatory gesture, but rather the acoustic onset. Our study on coordination between acoustic, articulatory and respiratory events prior to speech initiation shows a close coupling between exhalation onset and the articulatory gesture towards the initial segment. These findings suggest that respiration is not an automatic process providing the egressive airflow for speaking, but is integrated in the phonetic encoding of speech (Fuchs and Rochet-Capellan 2021).

5. References

Bates, D., M. Mächler, B. Bolker, and S. 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.

Browman, C. P. and L. Goldstein (1992). "Articulatory phonology: An overview". In: *Phonetica* 49.3-4, pp. 155–180.

Fuchs, S., C. Petrone, J. Krivokapić, and P. Hoole (2013). "Acoustic and respiratory evidence for utterance planning in German". In: *Journal of Phonetics* 41.1, pp. 29–47.

Fuchs, S. and A. Rochet-Capellan (2021). "The respiratory foundations of spoken language". In: *Annual Review of Linguistics* 7, pp. 13–30.

Gick, B., I. Wilson, K. Koch, and C. Cook (2004). "Language-specific articulatory settings: Evidence from inter-utterance rest position". In: *Phonetica* 61.4, pp. 220–233.

Krause, P. A. and A. H. Kawamoto (2021). "Predicting One's Turn With Both Body and Mind: Anticipatory Speech Postures During Dyadic Conversation". In: *Frontiers in Psychology* 12, p. 684248.

Krivokapić, J. (2014). "Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1658, p. 20130397.

Krivokapić, J., W. Styler, and B. Parrell (2020). "Pause postures: The relationship between articulation and cognitive processes during pauses". In: *Journal of Phonetics* 79, p. 100953.

Mooshammer, C., L. Goldstein, H. Nam, S. McClure, E. Saltzman, and M. Tiede (2012). "Bridging planning and execution: Temporal planning of syllables". In: *Journal of Phonetics* 40.3, pp. 374–389.

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

Ramanarayanan, V., E. Bresch, D. Byrd, L. Goldstein, and S. S. Narayanan (2009). "Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation". In: *Journal of the Acoustical Society of America* 126.5, EL160–EL165.

Rasskazova, O., C. Mooshammer, and S. Fuchs (2018). "Articulatory settings during inter-speech pauses". In: *Proceedings of the 13. Conference on Phonetics & Phonology in German-speaking Countries*, pp. 161–164.

— (2019). "Temporal Coordination of Articulatory and Respiratory Events Prior to Speech Initiation". In: *INTERSPEECH*, pp. 884–888.

Rastle, K., K. P. Croot, J. M. Harrington, and M. Coltheart (2005). "Characterizing the motor execution stage of speech production: consonantal effects on delayed naming latency and onset duration." In: *Journal of Experimental Psychology: Human Perception and Performance* 31.5, p. 1083.

Rochet-Capellan, A. and S. Fuchs (2014). "Take a breath and take the turn: how breathing meets turns in spontaneous dialogue". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1658, p. 20130399.

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

Schaeffler, S., J. M. Scobbie, and F. Schaeffler (2014). "Measuring reaction times: vocalisation vs. articulation". In: *Proceedings of the 10th ISSP*.

Slifka, J. (2003). "Respiratory constraints on speech production: Starting an utterance". In: *Journal of the Acoustical Society of America* 114.6, pp. 3343–3353.

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

Whalen, D. H. and J. M. Kinsella-Shaw (1997). "Exploring the relationship of inspiration duration to utterance duration". In: *Phonetica* 54.3-4, pp. 138–152.

Zöllner, A., C. Mooshammer, O. Rasskazova, and S. Fuchs (2021). "Breathing affects reaction time in simple and delayed naming tasks". In: *Proceedings of the 12th ISSP*, pp. 218–221.

Dialect specific patterns of gestural timing? Evidence from lateral clusters.

Emily Gorman[1]

[1]*Lancaster University*

e.gorman@lancaster.ac.uk

Abstract

This study explores the relationship between articulatory variation and speech timing, focussing on patterns of onset cluster timing in articulatorily distinct productions of /l/. Motivated by findings of variable cross-linguistic patterns of lateral cluster timing, this study compares lateral onset clusters across two closely related dialects of British English which differ in lateral darkness. Durational measures are used to determine the stable intervals and temporal movements across singleton and cluster pairs. Unexpectedly, the study finds no effects of lateral darkness on lateral onset cluster timing. Possible explanations for these results are explored.

Keywords: Speech Timing; Laterals; Consonant Clusters; C-centre; Articulatory Data

1. Introduction

What conditions cluster timing? Syllable structure and language are among the factors shown to influence patterns of cluster timing (e.g., Marin and Pouplier, 2010; Shaw et al., 2011). For other factors, such as the intrinsic properties of segments, their effects on timing remains more tentative. One such example is the proposed effect of lateral darkness on cluster timing (Marin and Pouplier, 2014). The lateral segment is one of considerable complexity and variation. Within British English dialects alone, acoustic and articulatory realisations of /l/ differ markedly (e.g., Turton, 2014). Such within language variation provides a test case for measuring the effects of lateral darkness on cluster timing.

1.1. Patterns of lateral cluster timing

The C-centre pattern, regarded the common timing pattern for onset clusters within branching languages, (Browman and Goldstein, 1988) describes the presence of a stable temporal relationship between the centre of the consonantal unit and the following vowel across singleton and cluster contexts, as illustrated in Figure 1. For this to occur, relative to the singleton context, C1 of the cluster must shift leftwards away from the vowel, and C2 must shift rightwards towards the vowel. For a C-centre pattern to emerge, the temporal shifts of C1 and C2 must be equal. Explanations for C-centre timing patterns have been offered, most notably by proponents of a coupled oscillator approach to speech timing (e.g., Browman, Goldstein, et al., 1995). From this perspective, a C-centre patterns arises due to competing phase relationships. Both C1 and C2 are coupled in-phase with the vowel; however, both consonants cannot be produced concurrently with the vowel. The anti-phase relationship between C1 and C2 thus facilitates a compromise solution whereby consonants shift equally around the consonant centre, thus preserving the global relationship between consonant and vowel segments.

For lateral clusters, previous studies have reported typical C-centre patterns for lateral onset clusters in American English (Browman and Goldstein, 1988; Marin and Pouplier, 2010), Romanian (Marin and Pouplier, 2014), and Italian (Hermes, Mücke, and Grice, 2013). Challenging these results, however, are findings for non C-centre patterns in lateral onset clusters in English (Goldstein et al., 2009), German (Brunner et al., 2014) and Montreal French (Tilsen et al., 2012). For example, in an analysis of English speakers, Goldstein et al. (2009) observed an asymmetrical shift pattern in lateral clusters, such that in a /p/ + /l/ + V structure, /l/ shifted less than /p/.

The timing pattern predicted for coda clusters of branching languages is a sequential or local pattern (Browman and Goldstein, 1988). Within this pattern, the transition from a singleton coda to a coda cluster involves the simple addition of a second consonant, with no temporal effect on the previous consonant. Again, explanations for the sequential timing patterns of codas can be gained from the coupled oscillator model of speech timing. From this perspective, a sequential timing pattern arises due to non-competitive anti-phase coupling relationships between segments. Findings for the timing patterns of lateral coda clusters, as with onsets, are varied. While a predictable sequential timing pattern has been observed for lateral coda clusters in German (Pouplier, 2012), non sequential timing patterns have been reported for lateral coda clusters in American English (Marin and Pouplier, 2010). For American English speakers, Marin and Pouplier (2010) found that /l/ shifted leftward towards the preceding vowel within a V + /l/ + C2 sequence, relative to its timing in a singleton context, (V + /l/).

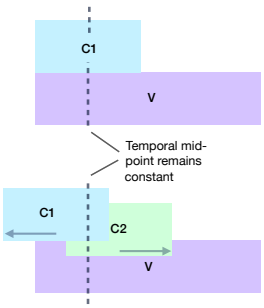


Figure 1: *Schematic illustration of C-centre timing pattern for onset cluster. A singleton, C1 + V, context is shown on top, and a cluster, C1 + C2 + V context is show on the bottom.*

The picture then for lateral cluster timing is varied. One explanation is that lateral cluster timing may be mediated by lateral darkness (Marin and Pouplier, 2010, 2014). Pursuing this hypothesis further, Marin and Pouplier (2014) investigated the timing patterns of liquids /l/ and /r/ in Romanian, which differ in

darkness. This aim of their study was to ascertain whether articulatory darkness affected timing patterns of liquid coda clusters. Indeed, results of their study were suggestive of such an interaction, with the dark coda /r/ of Romanian speakers patterning similarly to the dark coda /l/s of American English speakers (Marin and Pouplier, 2010).

While there is some cross-linguistic evidence that differences in lateral cluster timing can be explained by the articulatory composition of the lateral, the potential confounds presented by cross-linguistic evidence prevents any firm conclusions from being drawn. In response to this problem, this study explores lateral darkness in two closely related dialects of a single system (British English). While sharing the phonotactic constraints of a single system, dialects differ in articulatory patterns of lateral darkness. It is hoped that this design will facilitate an explicit test of the effects of lateral darkness on lateral cluster timing.

2. This Study

Lateral clusters were compared in two dialects of British English, namely Standard Southern British English (hereafter, SSBE), and Lancashire / Manchester English. Dialects were selected on the grounds of reported differences in lateral darkness between these dialects. While SSBE speakers are reported produce a clear /l/ in onset position, and a dark /l/ in coda position (e.g., Turton, 2014), Lancashire /Manchester speakers are reported to have dark /l/s in all positions (e.g., Hughes, Trudgill, and Watt, 2012).

3. Research Questions and Hypotheses:

Using dialect as a proxy for onset lateral darkness, the research question and hypotheses for this study are as follows:

RQ: How does lateral darkness interact with patterns of lateral onset cluster timing?

Hypotheses: Differences in lateral darkness will correlate with differences in lateral onset cluster timing. Specifically, clearer onset /l/ clusters (SSBE) are predicted to correlate with a non-C-centre timing pattern, given findings for non C-centre patterns in German where /l/ is relatively clear (Brunner et al., 2014). Darker onset /l/ clusters (Lancashire /Manchester) are rather predicted correlate with a C-centre timing pattern, given findings for C-centre patterns in American English where /l/ is relatively dark (Marin and Pouplier, 2010).

4. Method

Audio-synchronised electromagnetic articulography data was collected using the Carstens AG501 articulograph, recorded at 1250 Hz, and downsampled to 250 Hz. Audio data was recorded using a DPA 4006A microphone. Data was collected from 8 SSBE and 6 Lancashire / Manchester speakers. Acoustic and articulatory data was recorded while participants read sentences aloud from an adjacent screen. Sensors were attached mid-sagittally to approximately 1cm behind the tongue tip, the tongue dorsum (as far back as was comfortable for the participant), and the tongue body, which was positioned equidistant between the tip and the dorsum. Further sensors were attached mid-sagittally to the upper and lower lips and the gumline of the lower incisors. Reference sensors were also used to correct for head movements; these were attached to non-mobile structures including, the bridge of the nose, behind each ear, and the gumline of the upper incisors. Ear and nose sensors were secured to clear goggles which were worn by the participant throughout the experiment. Finally, a bite plate was use to rotate sensor movements to the occlusal plane.

Stimuli consisted of target words within the carrier phrase *"Say tea xx again"*. Target words contained /l/ within an onset cluster (/p l/ or /k l/) or a singleton context (/l/ + V), giving 4 cluster - singleton pairs. For each pair, vocalic context varied between front and back vowels, see Table 1. Each target word was repeated four times.

Cluster	Singleton
Plug	Lug
Plick	Lick
Club	Lug
Clip	Lip

Table 1: *Target token pairs.*

Acoustic segmentation was performed using Montreal Forced Aligner (McAuliffe et al., 2017) in Praat (Boersma, 2011). Gestural maxima for /p, b/, /k, g/ and /l/ were defined as the time point when the relevant displacement measure reached its velocity minimum. The relevant measure for /p, b/ was the lip aperture in the horizontal/vertical dimension, for /k, g/, it was the tongue dorsum displacement in the vertical dimension, and for /l/, it was tongue tip displacement in the verticl dimension. The velocity minima were identified automatically using a function for finding peaks in "pracma" package (Borchers, 2022). Checks for accuracy were performed, and adjustments were made to parameters such as the search window and minimum peak height where necessary.

Following a methodology adapted from Marin and Pouplier (2010), two sets of timing measures were calculated: (i) lateral to anchor lags, and (ii) stability timing measures. Lateral to anchor lags measured the duration between the target achievement of the lateral, and the target achievement of the post-vocalic consonant, or the "anchor" consonant. For example, in the *"plug / lug"* pair, the lateral to anchor lag was the time of target achievement of /g/ (the anchor consonant) minus the time of target achievement of /l/. Lateral to anchor lags were compared between singleton and cluster tokens of each word pair. Stability measures were then used to calculate the most stable interval across the singleton and cluster tokens of each word pair. Intervals included the duration of the C-centre to anchor for both the singleton and cluster tokens, and two additional measures for cluster tokens only. For singleton tokens, the C-centre was defined as the target achievement of the singleton consonant. For cluster tokens, the C-centre was defined as the temporal midpoint between the targets of C1 and C2. Cluster only lags included a left-edge to anchor lag (target of anchor to the target of C1), and a right edge to anchor lag (target of anchor to the target of C2). See Figure 2 for a schematic illustration of stability intervals in singleton / cluster pairs. Model comparisons were used to determine the degree of similarity between singleton and cluster intervals; details of the model structures are provided in Section 5.2.

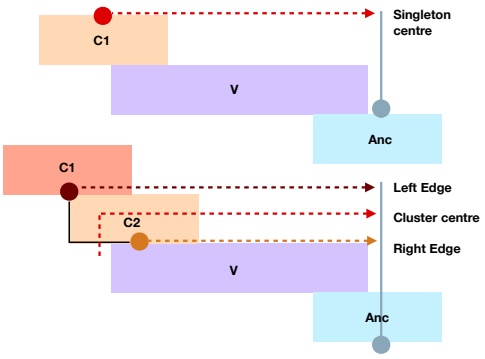


Figure 2: *Schematic illustration stability lag intervals.*

5. Results

5.1. Lateral to anchor lags

Figure 3 shows lateral to anchor lags for each singleton cluster pair, with dialect indicated by colour. A C-centre effect predicts the duration of the lateral to anchor lag will be shorter within the cluster context compared to the singleton context. This is because, C1 in a cluster must shift leftwards away from the vowel, and C2 must shift right towards the vowel. From Figure 3, we can see the the cluster context, on the left of each panel has a shorter lateral to anchor lag than the singleton context, on the right hand side of the panel. This is the case for each word pair, and for each dialect. While there is greater variation within the lag durations of SSBE, there are no qualitative differences between dialects.

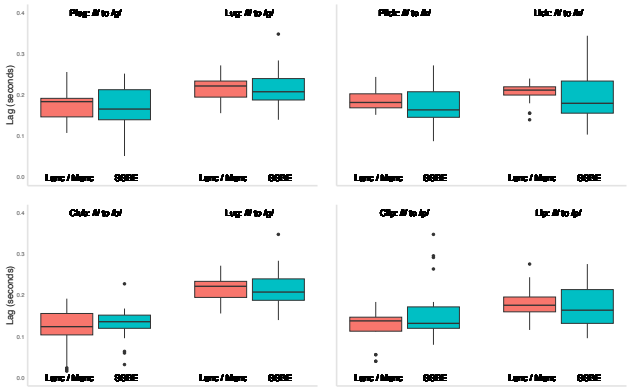


Figure 3: *Figure showing lateral to anchor lags in (s) for singleton - cluster pairs. Each panel shows a different pair.*

5.2. Stability measures

To determine the most stable interval across each singleton-cluster pair, and the effect of dialect on interval duration, linear mixed effects models were performed using the "lme4" package (Bates et al., 2015). For each word pair, three models were performed. Model (1) included the singleton centre lag and cluster centre lag; model (2) included the singleton centre lag and cluster left-edge lag; model (3) included the singleton centre lag and cluster right-edge lag. This structure enabled an explicit comparison between the singleton centre lag, and each of cluster lags. The cluster interval which was not significantly different to the singleton centre interval was considered the most stable interval across a word pair. Models included fixed effects of dialect and consonant structure (i.e., a term to test whether there was a significant difference between the duration of the singleton and consonant intervals included within the model), an interaction term between dialect and consonant structure, a random intercept of speaker, and a by-speaker random slope for consonant structure.

To test for the significance of dialect and consonant structure, model comparisons were performed using likelihood ratio tests. An effect was here considered significant if the model comparison was significant at $p < .05$. Full models were compared to partial models where an effect had been removed.

A C-centre pattern predicts a significant difference between models comparing the singleton centre lag and the cluster left-edge lag, and the singleton centre lag and the cluster right-edge lag. This is because across singleton and cluster pairs, these intervals are not held constant within a C-centre structure. Conversely, a C-centre pattern predicts a non significant difference between models comparing the singleton C-centre lag and the cluster C-centre lag, for these intervals are predicted to remain stable across singleton and cluster pairs.

For all word pairs, the effect of dialect on interval duration was non-significant, as was an interaction between dialect and consonant structure. However, each word pair differed regarding the interval of greatest stability. For *"plug / lug"*, a significant difference was found between the single centre interval and cluster left-edge lag ($p < .001$). This means that across the *"plug / lug"*, pair, the C-centre lag and the right-edge lags were both stable. For *"plick / lick"*, a significant difference was found between the singleton centre interval and all three cluster intervals, meaning that non of the intervals were stable across the singleton-cluster pair. For *"club / lug"*, a significant difference was found between the singleton centre lag and the cluster right edge lag only ($p = 0.016$), meaning that the C-centre and left-edge intervals were stable across the word pairs. Finally, for *"clip / lip"*, a significant difference was found between the singleton centre and the cluster left edge interval only ($p = 0.003$), meaning that the C-centre and right-edge intervals were stable across the word pair.

5.3. Results summary

Considering results from the lag measures and stability measures, no differences in lateral onset cluster timing were observed between dialects. In addition, stability measures showed that the C-centre was not typically the most stable interval across singleton and cluster pairs. The stability of the left-edge, right-edge, and C-centre intervals rather varied across word-pairs. For two of the word-pairs, *"plug / lug"* and *"clip / lip"*, the C-centre and right-edge were both stable. For *club / lug"* the C-centre and left-edge were both stable, while for *"plick / lick"* no interval was stable across the singleton/cluster word pair.

6. Discussion and conclusion

This study has examined the timing of lateral onset clusters across SSBE and Lancashire / Manchester dialects, where lateral darkness is reported to differ. Findings from the stability measures analysis showed that the C-centre was typically not

the most stable interval across singleton and cluster pairs for either dialect. Though this ran counter to the hypothesis that the dark /l/s of Lancashire / Manchester speakers would yield a C-centre pattern in lateral onset clusters, this result was not entirely surprising, given considerable variability reported for onset cluster timing patterns (Mücke, Hermes, and Tilsen, 2020).

A more surprising finding, was that neither the lag analysis nor the stability analysis show a timing difference between dialects. This result was unexpected given: (i) previous findings for lateral darkness and lateral cluster timing to show an apparent pattern of covariation, as discussed within the introductory section, and (ii) the centrality of timing to articulatory accounts of lateral darkness (e.g., Sproat and Fujimura, 1993). To confirm that the speakers within this study indeed differed in lateral darkness, an additional articulatory analysis was performed. Analysis showed a clear dialectal difference in lateral darkness. Secure in the knowledge of dialectal difference in lateral darkness, we are then faced with an interesting question: How can a stable timing pattern occur in lateral onset clusters which differ in lateral darkness? There are several avenues which could be explored in response to this question. I will here only speculate on a few.

The first possibility I consider is the presence of compensatory strategies which preserve timing while accommodating differences in lateral darkness. One such strategy is higher velocity. For example, higher velocity could enable a spatially larger dorsal gesture to be achieved without incurring a further temporal cost. Another possible compensatory strategy is a reduction in vowel duration. Since lag measures in this study span from a point within the consonant onset to the target of the post vocalic anchor consonant, systematic changes in vowel duration could reasonably influence timing measures.

Another factor could be the use of the tongue tip gesture to define lateral timing. Since the dialectal differences in lateral darkness manifest in differences in tongue body vertical displacement, it may be the case that the timing of the tongue tip gesture is not the most informative point within the lateral in terms of capturing the interaction between lateral darkness and cluster timing. The timing of the lateral tongue body gesture may be more informative in this regard; however, this is difficult to obtain within vocalic contexts.

In conclusion, this study has compared lateral onset cluster timing across two varieties of British English which differ in lateral darkness. That no timing differences were found across these varieties poses an interesting question regarding how temporal stability can be maintained across lateral clusters which differ in lateral darkness.

7. Acknowledgements

I would like to thank my supervisors, Dr Sam Kirkham, and Dr Patrycja Strycharczuk, for their considerable support and input on this piece of work. This work was supported by the Economic and Social Research Council; grant number: ES/P000665/1

8. 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.

Boersma Paul Weenink, David (2011). *Praat: doing phonetics by computer [Computer program]. Version 5.2.29*. http://www.praat.org/.

Borchers, Hans W. (2022). *pracma: Practical Numerical Math Functions*. R package version 2.4.2. https://CRAN.R-project.org/package=pracma.

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, Louis Goldstein, et al. (1995). "Dynamics and articulatory phonology". In: *Mind as motion: Explorations in the dynamics of cognition* 175, p. 194.

Brunner, Jana, Christian Geng, Stavroula Sotiropoulou, and Adamantios Gafos (2014). "Timing of German onset and word boundary clusters". In: *Laboratory Phonology* 5.4, pp. 403–454.

Goldstein, Louis, Hosung Nam, Elliot Saltzman, and Ioana Chitoran (2009). "Coupled oscillator planning model of speech timing and syllable structure". In: *Proceedings of the 8th Phonetic Conference of China and the International Symposium on Phonetic Frontiers in Phonetics and Speech Science*.

Hermes, Anne, Doris Mücke, and Martine Grice (2013). "Gestural coordination of Italian word-initial clusters: the case of 'impure s'". In: *Phonology* 30.1, pp. 1–25.

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.

— (2014). "Articulatory synergies in the temporal organization of liquid clusters in Romanian". In: *Journal of Phonetics* 42, pp. 24–36.

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.

Mücke, Doris, Anne Hermes, and Sam Tilsen (2020). "Incongruencies between phonological theory and phonetic measurement". In: *Phonology* 37.1, pp. 133–170.

Pouplier, Marianne (2012). "The gestural approach to syllable structure: Universal, language- and cluster-specific aspects." In: *Speech Planning and Dynamics*. Ed. by Susanne Fuchs, M Weihrich, D Pape, and D Perrier. Frankfurt am Main: Peter Lang, pp. 63–96.

Shaw, Jason A, Adamantios I Gafos, Philip Hoole, and Chakir Zeroual (2011). "Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters". In: *Phonology* 28.3, pp. 455–490.

Sproat, Richard and Osamu Fujimura (1993). "Allophonic variation in English /l/ and its implications for phonetic implementation". In: *Journal of Phonetics* 21.2, pp. 291–311.

Tilsen, Sam, Draga Zec, Christina Bjorndahl, Becky Butler, Marie-Josee L'Esperance, Alison Fisher, Linda Heimisdottir, Margaret Renwick, and Chelsea Sanker (2012). "A cross-linguistic investigation of articulatory coordination in word-initial consonant clusters". In: *Cornell Working Papers in Phonetics and Phonology* 2012, pp. 51–81.

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.

Dimensions of structure and variability in the human vocal tract

Katherine Vaughan-Williams[1], Steven Moran[2,3], Sam Kirkham[1]

[1]*Lancaster University, UK*
[2]*University of Neuchâtel, Switzerland*
[3]*University of Miami, USA*
kpvaughanwilliams@gmail.com, steven.moran@unine.ch, s.kirkham@lancaster.ac.uk

Abstract

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, but variation in one dimension is rarely independent of other dimensions. Are some of these relationships more variable than others, or do there exist invariants in the morphology of the vocal tract? In this study, we report a data-driven investigation into the relationship between vocal tract dimensions based on multi-speaker real-time magnetic resonance imaging data. We discover different sub-populations in the data, which correspond to groups of speakers that share a common relationship between vocal tract parameters. This suggests a range of complex patterns of co-variation in the morphology of the human vocal tract. We conclude by speculating on the possible implications of these results for understanding individual differences in speech production.

Keywords: vocal tract anatomy, magnetic resonance imaging, speaker-specific variation, conditional inference trees

1. Introduction

The human vocal tract exhibits considerable variation between speakers. Most obvious are the changes that accompany child development from birth until adulthood, whereby changes in vocal tract length are largely determined by growth in the pharyngeal regions (Vorperian et al. 2005). But we also observe variation in adult populations, 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). In many cases, variation in one dimension is rarely independent of other dimensions. For example, vocal tract dimensions are sometimes correlated with other aspects of the body, such as speaker height and weight (Stone et al. 2018), although in other cases there are no such relationships between speaker weight and vocal tract length (Hatano et al. 2012). In terms of variation within the vocal tract, the length of horizontal vocal tract structures tends to negatively correlate with the length of vertical structures (Honda et al. 1996), yet we also know that scaling is not uniform across different structures.

The fact that the relationship between vocal tract parameters varies between studies has many possible explanations, ranging from measurement technique to data quality to sample size. Aside from these considerations, one possible explanation is that the relationship between vocal tract parameters may be different in different areas of the parameter range. For example, speakers with a longer vocal tract may show a simple linear relationship with palate length, whereas perhaps speakers with a shorter vocal tract show a more complex relationship with palate length that could interact with other factors. This raises a question: 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? Such results have implications for patterns of variability in speech production, because anatomical differences place constraints on the use of particular speech production strategies (Fuchs, Winkler, and Perrier 2008; Brunner, Fuchs, and Perrier 2009; Weirich and Fuchs 2013). In order to address this question, we conduct an exploratory study into variation in the morphology of the human vocal tract, with the aim of understanding structured variability in the relationship between vocal tract dimensions using multi-speaker real-time magnetic resonance imaging data.

2. 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 two-dimensional midsagittal images of the vocal tract by the first author. All measurements were based on a single representative rest posture for each speaker and annotations were carried out using ImageJ (Schneider, Rasband, and Eliceiri 2012). The measurements reported in this study are as follows:

1. vocal tract length (mm)
2. palate length (mm)
3. palate height (mm)
4. tongue length (mm)
5. tongue area (mm^2)
6. body height (cm)
7. body weight (kg)
8. body-mass index (kg/m^2)

Our analysis is twofold: (1) what are the primary dimensions of variability? (2) what are the relationships between vocal tract parameters? We address (1) by submitting all measures to Principal Components Analysis, following by k-means clustering, which allows us to observe the ways in which measurements cluster together on a global scale.

The second analysis then aims to better understand the precise relationship between vocal tract parameters. A large number of highly-correlated measurements presents significant problems for modelling using classical parametric statistics, so we instead turn to a class of data-driven machine learning algorithms: conditional inference trees. Conditional inference trees are a class of regression models using binary recursive partitioning. We first test the null hypothesis of independence between

the outcome variable and each predictor variable. If the null hypothesis cannot be rejected then the process stops. If the null hypothesis can be rejected then we select the predictor variable that has the strongest association with the outcome variable. We then implement a binary split in the predictor variable that maximises the homogeneity of each group in the binary split, in terms of its relationship with the outcome variable. This process is then repeated recursively until some stopping criterion is achieved, such as a maximum tree depth or minimum node size. The resulting model is a hierarchical tree with the most important predictor at the top and a series of binary splits within this predictor, which continues until all significant predictors have been exhausted. We visualise the models as in Figure 4, 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.

We implement conditional inference trees in R using the `partykit` package (Hothorn and Zeileis 2015). We fitted a conditional inference tree to each variable in the data set as the outcome variable, with all remaining variables as predictor variables. All p-values for the splits were calculated using the Bonferroni method.

3. Results

3.1. PCA

We find that two principal components capture 79.5% of the variance. As shown in Figure 1, these dimensions capture variation across (1) vocal tract length and tongue length/area, and (2) 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 in Figure 2, which highly correlate with speaker sex.

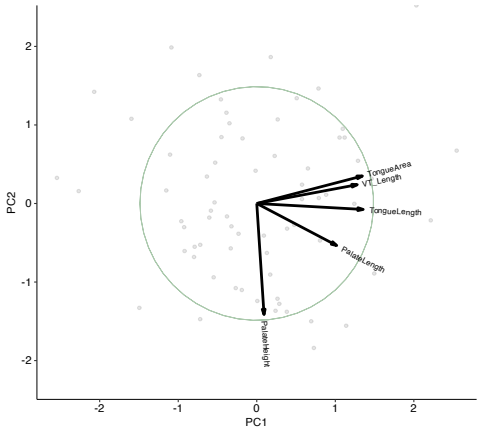


Figure 1: *PCA loadings for PC1 and PC1.*

3.2. Correlation matrix

Before showing the conditional inference trees, we first explore simple pairwise correlations between measurements. Figure 3 shows a correlation matrix for all variables. BMI is unsurprisingly highly correlated with height (r = 0.82), given that height

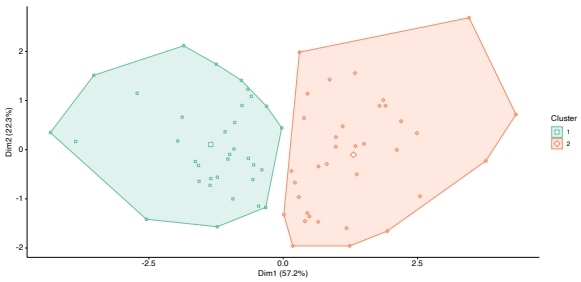


Figure 2: *Cluster plot showing each speaker in two-dimensional PCA space.*

is incorporated into the BMI measure. The next strongest correlation is between tongue area and tongue length (r = 0.79), which is also unsurprising given the inherent physical relationship between these measures. We also observe moderately strong correlations between tongue area and vocal tract length (r = 0.75), height and vocal tract length (r = 0.74), and tongue length and vocal tract length (r = 0.71). One problem with this analysis is that such variables are likely to be highly correlated with a number of other variables. Our following analysis addresses this point using conditional inference trees, which are well-suited to exposing complex relationships in highly colinear data.

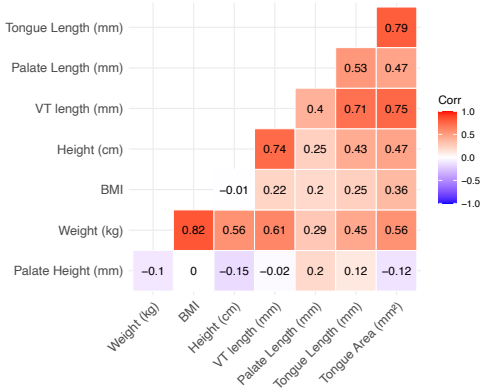


Figure 3: *Correlation matrix for all vocal tract and body measurements in the dataset.*

3.3. Conditional inference trees

The conditional inference trees expose more precise relationships between parameters. We show visualisations for three conditional inference trees that reveal the most interesting relationships and summarise some of the other results in text.

Figure 4 shows a conditional inference tree 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 the strongest predictor of vocal tract length, with male speakers having longer vocal tracts than female speakers. Within male speakers, there is one split in the distribution, such that speakers with a smaller tongue area (below or equal to 2937.6 mm^2) are more likely to have a smaller vocal tract. Within female speakers, a similar split occurs, but for tongue

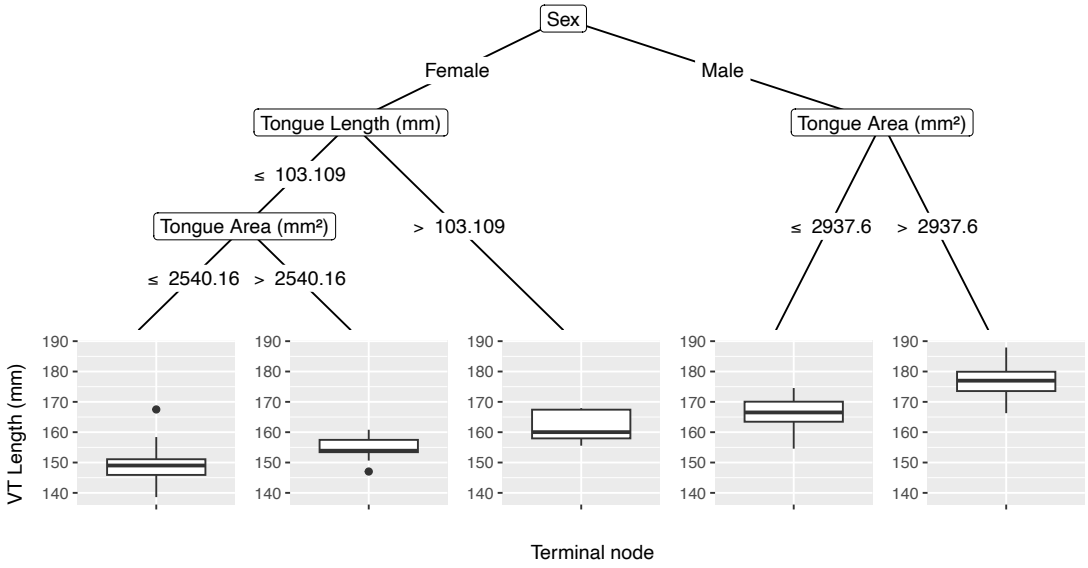


Figure 4: *Conditional inference tree fitted to vocal tract length measurements. Predictors that do not appear on the plot are not significant predictors of vocal tract length in the model.*

length rather than tongue area: speakers with longer tongues (greater than 103.109 mm) have longer vocal tracts. Finally, within female speakers with a shorter tongue, there is a further split based on small differences in tongue area, whereby a larger tongue area correlates with a slightly longer vocal tract. The other variables show no significant association with vocal tract length. This suggests a series of sub-populations in terms of how different measures impact vocal tract length in these data.

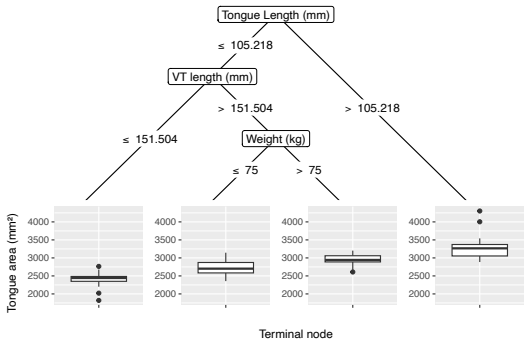


Figure 5: *Conditional inference tree fitted to tongue area measurements. Predictors that do not appear on the plot are not significant predictors of tongue area in the model.*

We fitted a conditional inference tree to tongue length, but found that variation in this measurement was only significantly predicted by variation in tongue area, which is unsurprising as we would expect a strong association between two related measures of the tongue. In the interests of space, we have not included a visualisation of this model. Instead, we show the model visualisation for the predictors of tongue area in Figure 5. This model shows that tongue length is the most important predictor of tongue area, with longer tongues predictably showing a larger area. However, within the lower half of the tongue length range (i.e. below or equal to 105.218 mm) other measures help to explain some of the variation. For example, in speakers with tongue length less than 105.28 mm there is an effect of vocal tract length in the expected direction. But in speakers with a slightly longer vocal tract there is a small difference in tongue area between speakers who weigh more than 75 kg and those that weigh less than 75 kg. This suggests that the relationship between tongue area and weight is a rather complex one and only emerges in a particular area of the range of possible tongue area values in these data. This is the only case where we found a significant relationship between a measurement of the whole body (such as height, weight, BMI) and a measurement of the vocal tract. In all other cases, none of the body measures were significant predictors of variation in vocal tract morphology.

Figure 6 shows a conditional inference tree fitted to palate length. In this case, the only variable that significantly predicts variation in palate length is tongue length. Specifically, speakers with a tongue length greater than 96.524 mm have a significantly longer palate than those with a tongue length below this value. The distributions between these two groups are fairly well separated, suggesting a strong association between tongue length and palate length.

4. Discussion and conclusion

We report a data-driven investigation into patterns of variability in the morphology of the human vocal tract. The most complex relationships are found in explaining the variance in vocal tract length. While the most important predictor is a fairly predictable sex-based difference, we then find that speakers with shorter vocal tracts also have smaller tongues (measured as tongue length in female speakers and tongue area in

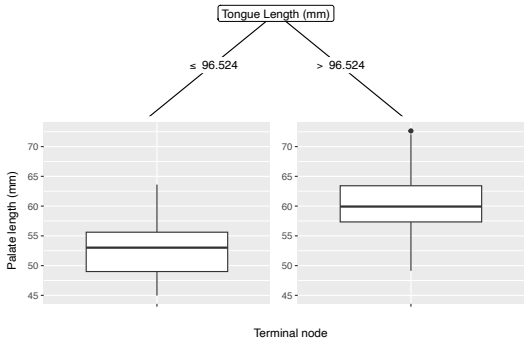


Figure 6: *Conditional inference tree fitted to palate length measurements. Predictors that do not appear on the plot are not significant predictors of palate length in the model.*

male speakers). Within female speakers, there is a sub-grouping of vocal tract length differences within speakers with smaller tongues, whereby those with smaller tongue areas have shorter vocal tracts. We note that these relationships are not uniform across speakers and point toward sub-groupings based on the interactions between vocal tract measurements.

Pairwise correlations showed moderately strong associations between vocal tract length and height, but we do not find this to be a significant predictor in our conditional inference trees, suggesting that this relationship can be captured via other dimensions in the model. In fact, we find relatively few relationships between vocal tract measurements and height/weight/BMI. The only significant effect of such a variable is in the model for tongue area, but the effect is limited. Specifically, the effect of weight on tongue area is only present for speakers with both a tongue length equal to or below 105.218 mm *and* a vocal tract length greater than 151.504 mm. Finally, we observed a simple relationship between tongue length and palate length, where speakers with longer tongues have predictably longer palates.

Overall, these results suggest that the relationship between vocal tract dimensions may vary across different sections of a parameter range, thereby complicating a straightforward scaling between dimensions. In terms of the implications of these results for speech production, it is unknown whether different sub-populations – as represented in the terminal nodes of our conditional inference trees – are likely to show any substantial differences in speech production. One possibility is that the anatomical constraints that characterise different subpopulations could lead to slight differences in articulatory behaviour. Whether such articulatory behaviours are motor equivalent and hence produce similar acoustic outputs is a possibility, but it is also worth investigating whether such anatomical differences underpin any of the observed individual variability in speech. Indeed, this raises the possibility that there could exist different classes of individual speaker variability that correspond with some of the sub-populations reported here.

In summary, this study reports the existence of subpopulations that share a set of relationships between vocal tract dimensions in different regions of the relevant parameter ranges. Future research will investigate whether individual variability in speech production can be grouped into similar classes that correspond to clusters of anatomical variation.

5. Acknowledgements

SK was funded by AHRC grant AH/Y002822/1, and SM was funded by SNSF grant PCEFP1_186841. Thanks to Tiena Danner.

6. References

Brunner, Jana, Susanne Fuchs, and Pascal Perrier (2009). "On the relationship between palate shape and articulatory behavior". In: *Journal of the Acoustical Society of America* 125.6, pp. 3936–3949.

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.

Fuchs, Susanne, Ralf Winkler, and Pascal Perrier (2008). "Do speakers' vocal tract geometries shape their articulatory vowel space?" In: *Proceedings of the International Seminar on Speech Production*, pp. 333–336.

Hatano, Hiroaki, Tatsuya Kitamura, Hironori Takemoto, Parham Mokhtaru, Kiyoshi Honda, and Shinobu Masaki (2012). "Correlation between vocal tract length, body height, formant frequencies, and pitch frequency for the five Japanese vowels uttered by fifteen male speakers". In: *Proceedings of Interspeech*, pp. 402–405.

Honda, Kiyoshi, Shinji Maeda, Michiko Hashi, Jim S. Dembowski, and John R. Westbury (1996). "Human palate and related structures: their articulatory consequences". In: *Proceedings of ICSLP '96* 2, pp. 784–787.

Hothorn, Torsten and Achim Zeileis (2015). "partykit: A modular toolkit for recursive partitioning in R". In: *The Journal of Machine Learning Research* 16.1, pp. 3905–3909.

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, Asterios Toutios, Yannick Bliesener, Ye Tian, Sajan Goud Lingala, Colin Vaz, Tanner Sorensen, Miran Oh, Sarah Harper, Weiyi Chen, Yoonjeong Lee, Johannes Töger, Mairym Lloréns Monteserin, Caitlin Smith, Bianca Godinez, Louis Goldstein, Dani Byrd, Krishna S. Nayak, and Shrikanth S. Narayanan (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.

Vorperian, Houri K., Ray D. Kent, Mary J. Lindstrom, Cliff M. Kalina, Lindell R. Gentry, and Brian S. Yandell (2005). "Development of vocal tract length during early childhood: a magnetic resonance imaging study". In: *Journal of the Acoustical Society of America* 117.1, pp. 338–350.

Weirich, Melanie and Susanne Fuchs (2013). "Palatal morphology can influence speaker-specific realizations of phonemic contrasts". In: *Journal of Speech, Language, and Hearing Research* 56, S1894–S1908.

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, Germany*

[2]*Laboratoire de Linguistique Formelle, CNRS, UMR 7110, Université Paris Cité, France*

[3]*Department of Rehabilitation and Special Education, University of Cologne, Germany*

jmertz1@uni-koeln.de, lena.pagel@uni-koeln.de, pperniss@uni-koeln.de,
giuseppina.turco@cnrs.fr, doris.muecke@uni-koeln.de

Abstract

Speakers tend to modulate the amount of coarticulatory cues according to the communicative needs at hand. Coarticulation has also been observed in the visual-gestural modality. Despite this, little is known about the use of coarticulatory strategies in sign language, probably partly because access to this fine-grained information can be very challenging. While the use of 3D electromagnetic articulography (EMA), a highly sophisticated experimental technique, has been widely (and successfully) tested on speech, the present studies are the first to provide precise kinematic measurements in the production of one deaf signer. In Study 1, we recorded signs produced in various locations on the signer's body in different setups (height of articulograph, hands' resting position, signer's position). After having identified locations and setups that can be used to optimize testing of sign kinematics, we ran an experiment on coarticulation (Study 2) in French Sign Language (LSF) to capture articulatory overlap between signs occurring next to each other. In this novel approach, we recorded a deaf native signer (EMA/video) while signing phonological pairs composed of '1'- and/or '3'-handshape. In a dynamical framework, we examine the kinematics of our sign data, revealing systematic patterns of overlapping organization driven by the phonological system. Our preliminary data showed both temporal and spatial dimensions of coarticulation in signing: (1) The anticipation of the '3'-handshape before the end of its immediately preceding '1'-handshape sign (and vice versa); (2) the truncation of the repetitive movement of the sign. Our findings speak in favor of the acquisition of kinematic data for capturing contextual variation phenomena.

Keywords: kinematics, sign language, coarticulation, electromagnetic articulography, methodology

1. 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 communicative demands 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 SL (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.

Unfortunately, these technologies show several limitations. In RF-sensing, for example, facial expressions cannot be captured, which is problematic since they can carry lexical and prosodic information, and minimal pairs can be found based on mouthing and mouth gestures (Crasborn, van der Kooij, et al. 2008). Motion capture, on the other hand, allows recording all body movements with no space restrictions when coupled with video recordings, but only a few linguistic laboratories use these devices.

An interesting and widespread technical alternative to motion capture is 3D electromagnetic articulography (EMA), as it offers a more cost-effective solution, potentially halving expenses. In spoken language, the use of EMA has been proven very effective to measure speech kinematics during coarticulation. EMA enables recordings of oral articulators such as tongue and lips movements in real time with high spatial and temporal resolution. This system provides precise 5-dimensional coordinates for each sensor position and it can trace changes over time throughout the sensors' movements. Additionally, EMA sensors can be fixed to both manual and vocal articulators, facilitating the study of mouthing in signing as well as bimodality (e.g., code-blending in hearing signers). It also allows for a unified experimental approach for both speakers and signers, emphasizing EMA's potential as an alternative method for studying sign kinematics. So far, EMA has not been used to study SL kinematics. The goal of the current study is to extend its use to SL and the fine analysis of the articulators used in the visual-gestural modality (i.e. the hands, head, torso, etc.) in contexts that are prone to undergo gestural overlap (coarticulation) between competing sublexical units.

We conducted two studies. First, we ran a methodological study (Study 1) to identify the conditions in which the use of EMA in SL is possible or not. The electromagnetic field generated by the articulograph is limited in terms of range for sensor detection. The goal was therefore to establish these limits to ensure correct detection of sensors placed on the articulators of SL with EMA. Thus, identifying the maximum distances between the sensors and emitting coils will allow control over the parameters to be tested and the categories of signs when designing future studies. Second, we conducted an experiment (Study 2) to test coarticulation in both handshape and location in French

Sign Language (LSF) as a case study for the use of EMA in SL. Study 2 is the first study on coarticulation in LSF using EMA to quantify a signer's gestural behavior (i.e. *articulatory* gestures). Taken together, these two studies aim to identify the methodological adjustments required to build a large study on SL using EMA. In the current paper, we present the results of a qualitative analysis of the data, and we focus on coarticulation in handshape.

2. Study 1: Methodological adjustments for the use of EMA in sign language

2.1. Participant

One deaf signer participated in the study. He was 34 years old, is right-handed, and a native signer of LSF, i.e. he was born deaf with two deaf signing parents.

2.2. Materials

The deaf participant was asked to produce individual signs in LSF distributed across four lists: signs produced (1) low on the torso (e.g., MAMAN "mother"), (2) far from the body (e.g., ALLER "to go"), (3) on the head (e.g., CONNAÎTRE "to know'), and (4) above the head (e.g., CHEF "head chef"). These four types of signs served to test the detection of the sensors within the magnetic field from various distances. Each list of signs was evaluated by combining three parameters: the height of the EMA (high vs. low EMA), the signer's position (standing vs. sitting), and the resting position of the hands (on the laps/side vs. on the belly vs. on a high table).

2.3. Procedure

The participant was provided with an information letter and consent form, both presented in written French and in LSF video recordings assured by the first author of the study, a hearing individual proficient in LSF. A video camera was positioned on the right diagonal of the signer to capture the upper body (including torso, hands, and face), which was subsequently used for video annotation and visual inspection of recordings. To monitor articulator movements, 15 EMA sensors were attached to the signer's head (forehead, behind ears), torso (sternum, shoulders), arms (wrists, mid-forearms), and right hand (thumb, index, middle and pinky fingernails, palm). A visual is provided in Figure 1.

Prior to the recording of each list, the participant saw and repeated all signs to confirm the accuracy of the lexical entries to be tested. Each list of signs was evaluated in each condition that corresponded to a separate EMA recording. To ensure a total reset of each individual sign, the participant returned his hands to the designated resting position, which varied based on the condition. The procedure was thoroughly explained to the participant before starting the experiment.

2.4. Data collection and analysis

The kinematic recordings were performed using 3D EMA (Carstens AG501) and a time-synchronized video setup. We used the EMA software (Carstens Medizinelektronik) to process the kinematic data, and the ema2wav converter (Buech et al. 2022) for the post-processing of each sensor. This yielded the 3D movement data of each sensor in terms of position, velocity and acceleration. The EMA data were recorded with 1,250Hz, downsampled to 250Hz and smoothed with a 3-step floating mean. For the video recordings, we used 50 frames per seconds. We used a clapperboard as auditory input for time-synchronization of EMA and video data.

Figure 1: *The sensors were taped to various body parts of the signer. This example illustrates the condition characterized by high EMA, the signer seated, with the high table for the hands.*

2.5. Results and interim discussion

The study aimed at assessing the efficacy of EMA in studying SL kinematics under various conditions. Taken together, findings revealed critical insights. Signs executed above the head were incompatible with low EMA, the hands were too close from the emitting coils which broke the signal. Additionally, when the signer placed his hands on his laps while seated or at his sides while standing in the resting position, signal detection was compromised, resulting in errors or undetected data. As a result, the use of a high table for hand resting positions proved to be a viable solution. In sum, if certain adjustments (EMA level, posture, etc.) are ensured, EMA can be an effective tool for studying SL kinematics.

The next steps consist in checking each sensor individually for signal breaks within each recording. Moving forward, forthcoming analyses will explore the correct detection of sensor orientation by the EMA device and delve into the temporal realization of SL compounds. Furthermore, two sweeps of natural signing will be analyzed, along with investigations into mouthing and mouth gestures during signing. These future analyses promise to enrich our understanding of SL articulation and sign dynamics, paving the way for more comprehensive studies in the field.

3. Study 2: Coarticulation in LSF

3.1. Participant

The participant was the same deaf signer who took part in the methodological study (cf. Study 1 above).

3.2. Materials

During the EMA recording, the signer was facing a computer monitor displaying the pairs of signs in the form of images. The

experiment was built using the software OpenSesame (Mathôt, Schreij, and Theeuwes 2012). 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 extension of the index finger (GERMAN, ORDER, HAVE-TO) and '3'-handshape to the extension of the thumb, index and middle fingers (ROOSTER, BAR, APARTMENT). To capture finger extension/closing, sign combinations included target pairs with X1 having the '1'-handshape 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, for a total of 66 trials. Moreover, the pairs were produced under 4 conditions in a block-wise fashion: (1) normal signing (i.e. *habitual* signing rate and speech register), (2) fast signing, (3) whispering, and (4) L2-directed speech, resulting in a total of 264 recorded trials. In the current paper we will report data from the normal signing condition.

3.3. Procedure

Based on the findings of Study 1, EMA sensors were placed on the signer's head, torso, arms and fingers. The session started with a training phase allowing the signer to familiarize with the task and to ensure a correct matching between the image and the sign. After that, the signer was asked to sign first under the normal condition so as to minimize a potential bias from the other rate/register conditions. There was a pause of about 5 minutes between the recording of each condition block. The whole session took about 2 hours.

We used ELAN (Crasborn and Sloetjes 2008) for video annotation and signal alignment of EMA transformed data in each trial. The data were analyzed in relation to 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). This framework allows for quantification of coarticulatory patterns, e.g., with respect to different speaking styles or communicative demands, and we aim to extend it to SL.

3.4. Data collection and analysis

The first step in the data processing was to define the location of the articulatory landmarks tracking the kinematic offset of X1 ("end of X1" from now onwards) and the onset of X2 ("beginning of X2"). The landmark criteria were determined based on the internal phonological movement of each individual sign. For example, the end of the sign HAVE-TO was defined based on the position of the wrist on the vertical axis (high-low position on the y-axis), while the beginning of the sign ROOSTER was defined based on the distance between the wrist and the forehead on the horizontal axis (front-back position on the x-axis; Figure 2). The kinematic landmarks served as delimiters of the X1 and X2 signs and were used as reference points for our following analyses of coarticulation in handshape.

The second step was to compute the transitional movements between the two signs. To do so, 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 (see Figure 2) were then detected automatically using a code in R (Team 2021). Quantitative analyses are currently in progress to strengthen the present descriptive analysis.

3.5. Results and interim discussion

The descriptive analysis of our articulatory data provides first kinematic evidence of the gestural overlapping of handshape movements at both temporal and spatial levels.

At a temporal level, we can observe anticipatory movements of handshape change before the end of X1 in many of the tested trials, including various signs in both '1-3' and '3-1' combinations. An example of extension/closing before the end of X1 is provided in Figure 2 below. As indicated by the green arrow, in the sequence HAVE-TO (sign composed of '1'-handshape) - ROOSTER ('3'-handshape) our signer anticipates the onset of the finger extension involved in the production of X2 before the end of X1.

The spatial dimension of coarticulation in handshape is attested by signs with internal organization that is phonologically composed of a repetitive movement (i.e. ROOSTER, GERMAN, APARTMENT, BAR). Crucially, the kinematic data allows detecting the partial-to-full truncation of the repetitive movement in a gradient way whereas this is not always visible through a frame-by-frame analysis based on video data.

4. General discussion

The use of 3D EMA in SL research has proven to be highly effective, enabling precise kinematic measurements and analysis within a dynamical framework. Our preliminary exploration of EMA setups 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). Because of phonetic and phonological constraints, signs, unlike (co-speech) gestures, are never produced very low, very high, or very far, which allows for relatively good detection of sensors placed on the distal parts of the body in the magnetic field. We observed that certain configurations of sign distances (low, far, high) and setups (e.g., high EMA, use or non-use of a table to rest the hands) are less suitable for the recording of movements of the articulators in SL. Study 1 has allowed us to carefully select parameters for Study 2 on coarticulation. Furthermore, it enables us to optimize the recording of longer sequences of signs (e.g., signed discourse) in future studies. Thanks to such methodological adjustments, we ran the first experiment on coarticulation in LSF (Study 2). The analysis of sign minimal pairs produced by one deaf native signer revealed systematic coarticulatory effects as well as intra-individual variations in normal signing. This first case study shows great promises for the use of EMA in the study of sign kinematics.

Additionally, 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 domains such as co-speech gestures and non-manual components in SL lexicons. This methodological approach extends its utility to bilingual bimodal speech, integrating the communicative role of visual cues for communication purposes and even the analysis of mouthing.

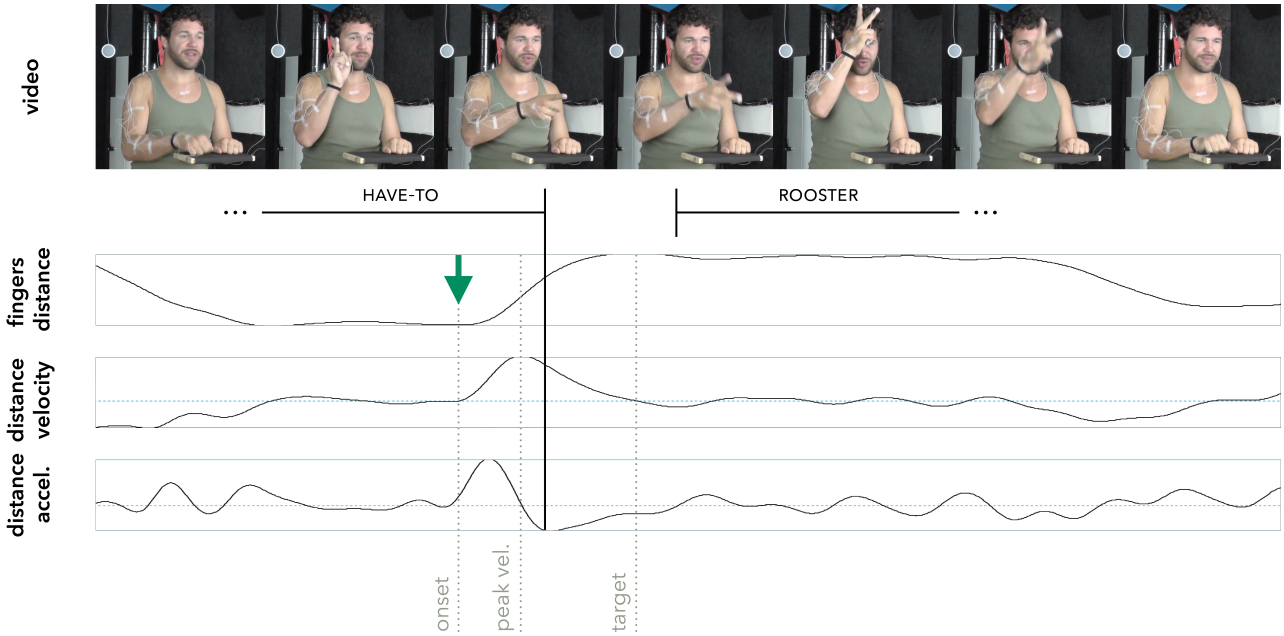


Figure 2: Example of coarticulation in the sequence HAVE-TO - ROOSTER ('1-3'). 3D Euclidean distance between the thumb and the pinkie finger, velocity and acceleration of extension show that the onset starts before the end of HAVE-TO.

5. Conclusion

Using EMA for SL kinematics exhibits substantial benefits. By implementing appropriate methodological refinements, its integration promises significant advancement in the field of SL phonology and phonetics.

6. Acknowledgments

We warmly thank Theodor Klinker for his invaluable work in building the experimental setup, conducting the EMA recordings, and processing the kinematic data, and Gyong Min Oh for taking pictures during the recordings. We sincerely thank our participant Thomas Lévêque for his participation in our studies. This work was supported by the Fyssen Foundation, the Laboratoire de Linguistique Formelle, the French Investissements d'Avenir-Labex EFL program (ANR-10-LABX-0083) contributing to the IdEx Université Paris Cité (ANR-18-IDEX-0001), the Cologne Center of Language Sciences (CCLS) and the German Research Foundation (DFG) as part of the SFB1252 "Prominence in Language" (Project-ID 281511265), project A04 "Dynamic modelling of prosodic prominence" at the University of Cologne.

7. References

Browman, C. P. and L. Goldstein (1992). "Articulatory phonology: An overview". In: *Phonetica* 49.3-4, pp. 155–180.

Buech, P., S. Roessig, L. Pagel, D. Muecke, and A. Hermes (2022). "ema2wav: doing articulation by Praat". In: *Interspeech*. Incheon, South Korea, pp. 1352–1356.

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: *Construction and Exploitation of Sign Language Corpora*, pp. 39–43.

Crasborn, O., E. van der Kooij, D. Waters, B. Woll, and J. Mesch (Jan. 2008). "Frequency distribution and spreading behavior of different types of mouth actions in three sign languages". en. In: *Sign Language & Linguistics* 11.1, pp. 45–67.

Gafos, A. I. and S. Benus (2006). "Dynamics of Phonological Cognition". In: *Cognitive Science* 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 W. J. Hardcastle and A. Marchal. NATO ASI Series. Dordrecht: Springer Netherlands, pp. 403–439.

Mathôt, S., D. Schreij, and J. Theeuwes (2012). "OpenSesame: An open-source, graphical experiment builder for the social sciences". In: *Behavior research methods* 44.2, pp. 314–324.

Mauk, C. E., B. Lindblom, and R. P. Meier (2008). "Undershoot of ASL locations in fast signing". In: *Signs of the time. Selected papers from TISLR* 8, pp. 3–24.

Mücke, D., A. Hermes, and T. Cho (Sept. 2017). "Mechanisms of regulation in speech: Linguistic structure and physical control system". In: *Journal of Phonetics*. Mechanisms of regulation in speech 64, pp. 1–7.

Team, R Core (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tyrone, M. E. and C. E. Mauk (2010). "Sign lowering and phonetic reduction in American Sign Language". In: *Journal of Phonetics* 38.2, pp. 317–328.

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

Sireemas Maspong, Francesco Burroni

Spoken Language Processing Group, Institute for Phonetics and Speech Processing, LMU München
s.maspong@phonetik.uni-muenchen.de, francesco.burroni@phonetik.uni-muenchen.de

Abstract

In Thai, vowel length contrasts have been reported to be primarily distinguished by differences in acoustic duration, with limited differences in vowel quality. Little is known about the articulatory characteristics of vowel length contrasts. In this paper, we 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 durational characteristics associated with the production of long vowels, compared to their short counterparts. These findings challenge conventional assumptions regarding Thai vowel length, which typically assume that long vowels are short vowels associated with longer durations, thus our accounts also question the target undershoot model.

Keywords: vowel length, articulatory, Thai

1. Introduction

In Thai, all monophthongs contrast for vowel length. Previous research has demonstrated that duration is the primary property of length contrast for all vowel pairs (Abramson 1962; Abramson and Ren 1990; Roengpitya 2001; Luangthongkam 2011; Onsuwan 2005): long vowels have longer acoustic duration than their short counterparts. It is worth noting that the acoustic duration of vowels in these studies was measured from the onset to the offset of the vowel voicing. Furthermore, perception studies confirmed that duration is the main cue to distinguish vowel length contrast in Thai (Abramson and Ren 1990; Roengpitya 2001). Studies have also indicated some differences in quality between short and long vowels in Thai, with short vowels tending to be more centralized in the vowel space, based on acoustic evidence (Abramson 1962; Luangthongkam 2011). Vowel quality has also been reported to play a "secondary" role in the perception of vowel length contrast in Thai as well (Abramson and Ren 1990; Roengpitya 2001).

Despite the substantial body of acoustic research dedicated to vowel length in Thai, our understanding of the articulatory characteristics of vowel length contrasts in Thai, particularly with regard to spatial and durational kinematic properties, remains limited.

Cross-linguistically, studies indicate a close association between vowel length contrasts and tense/lax contrasts. Long and short vowels have been observed to differ in quality, with short vowels often being more lax or centralized both acoustically and articulatorily (e.g., Lindblom 1963; Hoole and Mooshammer 2002; Harrington, Hoole, and Reubold 2012; Ratko, Proctor, and Cox 2023). One proposed explanation for the relationship between vowel length and quality comes from Lindblom (1963): short vowels are more centralized than long vowels due to limitations in their duration triggering target undershoot. In essence, long and short vowels share the same target, but because short vowels have a shorter duration to reach the target, they may undershoot. Consequently, the realization of short vowels tends to be more centralized than long vowels, which have a longer duration to reach the articulatory target. In some languages, vowel quality is independently manipulated from duration (*cf.* Ratko, Proctor, and Cox 2023 for a comprehensive literature review). Some interpretations suggest that this behavior represents a reanalysis of secondary cues (Garrett and Johnson 2013). Specifically, vowel quality, initially a by-product of durational differences, is reanalyzed as a primary cue. However, the target undershoot model has been challenged in several studies (e.g., Van Son and Pols 1990; Ratko, Proctor, and Cox 2023).

In this paper, we investigate the vowel length contrast in Thai. We chose Thai as a case study because, unlike in other languages previously studied articulatorily, duration has been consistently reported as the primary, if not sole, cue to vowel length contrasts in the language. In other words, Thai vowel length represents a more "pure" quantity contrast compared to languages like English or German, this is because Thai does not contrast tense and lax vowels. We focus specifically on the articulatory features of long and short vowels in Thai, leveraging results obtained from electromagnetic articulography (EMA). Our findings reveal distinct spatial and durational 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. The wider implication of our finding is that the difference between long and short vowels, even in a pure durational contrast like that of Thai, do not seem to be the result of undershooting at shorter durations.

1.1. Research questions

In this paper, we aimed to address two research questions concerning Thai vowel length:

1. Are short and long vowels in Thai distinguishable solely by their duration, or do they also exhibit differences in their articulatory properties?
2. If there are articulatory differences, are these differences derived from the undershoot of short vowels?

If the Thai vowel length contrast is indeed a "pure" quantity contrast, any observed articulatory disparities between short and long vowels may potentially stem from longer duration, as posited by the target undershoot model (Lindblom 1963).

2. Articulatory properties of vowel length

In this section, we investigated articulatory properties of short and long vowels, both static and dynamic properties, following Burroni et al. (2024, this volume).

2.1. Methods

Data were collected from a total of 6 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(:)/ or /i(:)/ 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(:)/ were surrounded by words with /i:/ vowel, while words with /i(:)/ were surrounded by words with /a:/ vowel. The choice of bilabial onset and coda was deliberate to minimize conflicting demands on tongue movement.

Participants were instructed to produce the stimuli at three different speech rates, thereby introducing variability in vowel duration. Vocalic gestures were identified by tracking tongue and jaw movements.

Landmarking of tongue movement was executed on the vertical displacement of the tongue body sensor, while landmarking of jaw movement was conducted based on the vertical displacement of the jaw sensor. Examples of tongue and jaw landmarks are shown in Figure 1. From the landmarking process, five measurements were extracted for each articulatory trajectory: (i) maximum tongue body height for vowel /i(:)/ and minimum jaw height for vowel /a(:)/, (ii) duration of the articulatory steady state of jaw and tongue body (the duration between articulatory target and release landmarks), (iii) movement amplitudes, (iv) peak velocity from onset to target, and (v) stiffness (calculated as the ratio of peak velocity and movement amplitude). The decision of analyzing tongue movement for /i(:)/ and jaw movement for /a(:)/ is based on the assumption that high vowels are produced with a more active tongue control, while low vowels are produced with a more active jaw control (see Mooshammer, Hoole, and Geumann 2007 for discussions on jaw control).

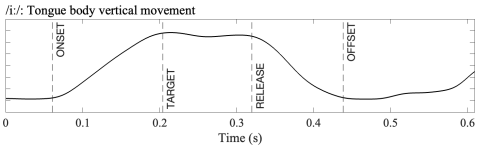


Figure 1: Landmarks of the tongue body vertical movement from an /i:/ token.

Each measurement was z-scored by participant, and linear mixed-effect regressions were fitted, treating each measurement as a dependent variable. Fixed effects included vowel length (long or short), utterance duration (z-scored), and their interaction. Utterance duration was calculated from the duration from the acoustic onset of a word preceding the target word to the acoustic offset of a word following the target and normalized by subtracting the target vowel duration, following the methodology of Tilsen and Tiede (2023). We introduced utterance into the model to capture the potential variations due to speech rate. Additionally, subject was included as a random intercept in the analysis.

2.2. Results

Our findings indicate distinctive articulatory patterns associated with long vowels, characterized by systematically more prominent movements, longer articulatory steady state, movement amplitude (for only jaw movement of vowel /a(:)/), and stiffness (for only jaw movement of vowel /a(:)/). This section presents the statistical results regarding the different kinematic properties of short and long vowels.

2.2.1. Tongue height and jaw height

For the vowel /i(:)/, our findings demonstrate that long vowels exhibit significantly higher maximum tongue height compared to their short counterparts ($t(327) = 4.78, p < 0.001$). The effect size is estimated at approximately 0.50 z-scores, indicating that, at the average utterance duration, long vowels have a higher maximum tongue height than short vowels by 0.50 z-scores. Similarly, for the long vowel /a:/, we observe a similar pattern: long vowels display significantly lower minimum jaw height than their short vowel counterparts ($t(336) = -7.43, p < 0.001$). The effect size is estimated at approximately -0.75 z-scores. See Figure 2.

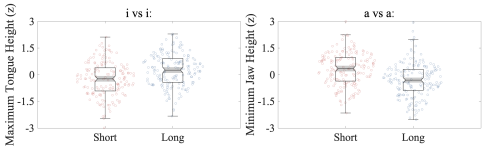


Figure 2: Maximum tongue height of vowel /i(:)/ (left) and minimum jaw height of vowel /a(:)/.

For the vowel /i(:)/, we found that the maximum tongue height is not influenced by utterance duration. There was no significant effect observed for utterance duration or the interaction between vowel length and utterance duration, indicating that the maximum tongue height of vowel /i(:)/ remains stable regardless of speech rate. However, for the vowel /a(:)/, we observed a significant effect of utterance duration on minimum jaw height ($t(336) = -3.49, p < 0.001$), with an effect size estimated at -0.21 z-scores. Nevertheless, we did not observe any significant effect of the interaction, suggesting that although the minimum jaw height is affected by speech rate, the difference in minimum jaw height between short and long vowels remains consistent.

2.2.2. Duration of articulatory steady state

Both vowel /i(:)/ and vowel /a(:)/ exhibit a similar pattern regarding the articulatory steady state duration of tongue and jaw movement, respectively. Long vowels demonstrate significantly longer steady state durations than their short counterparts (for /i/: $t(327) = 16.10, p < 0.001$; for /a/: $t(336) = 18.09, p < 0.001$). The effect sizes are estimated at around 1.25 and 1.39 z-scores, respectively. See Figure 3.

For vowel /i(:)/, we also observed significant effects of utterance duration ($t(327) = 2.63, p = 0.009$) and the interaction of vowel length and utterance duration ($t(327) = 4.73, p < 0.001$). The effect sizes are 0.15 and 0.37, respectively, indicating that the steady state duration of short vowel /i/ increases by 0.15 z-scores and the steady state duration of long vowel /i:/ increases by 0.52 z-scores when the utterance duration increases by 1 z-score. For vowel /a(:)/, we only observed a significant effect of the interaction ($t(336) = 8.70, p < 0.001$)

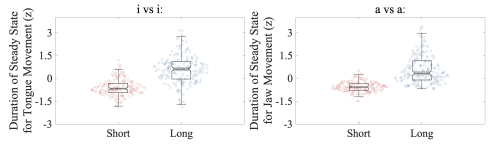


Figure 3: Steady state duration of tongue movement for vowel /i(ː)/ (left) and of jaw movement for vowel /a(ː)/.

with the effect size of 0.73 z-scores, indicating that only long vowels exhibit significant increased in steady state duration when the utterance duration increases.

2.2.3. *Movement amplitude*

We only detected a significant effect of vowel length on movement amplitude in jaw movement for vowel /a(ː)/ ($t(336) = 8.49, p < 0.001, Est. = 0.85$ z-scores), whereas no significant effect was observed on tongue movement for vowel /i(ː)/. See Figure 4.

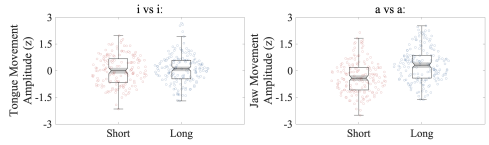


Figure 4: Amplitude of tongue movement for vowel /i(ː)/ (left) and of jaw movement for vowel /a(ː)/.

On the other hand, both tongue movement of vowel /i(ː)/ and jaw movement of vowel /a(ː)/ demonstrates a significant increase in movement amplitude when the utterance duration increases (for /i/: $t(327) = 3.71, p < 0.001, Est. = 0.31$ z-scores; for /a/: $t(336) = 3.47, p < 0.001, Est. = 0.21$ z-scores). We did not observe any significant effect of the interaction.

2.2.4. *Peak velocity from onset to target*

The only significant effect observed on the peak velocity from onset to target is the effect of utterance duration for tongue movement of vowel /i(ː)/ ($t(327) = -3.97, p < 0.001, Est. = -0.33$ z-scores). The negative effect size indicates that the peak velocity decreases with slower speech rates (increased utterance duration).

2.2.5. *Stiffness*

We only observed a significant effect of vowel length on stiffness in jaw movement for vowel /a(ː)/ ($t(333) = -11.55, p < 0.001, Est. = -1.07$ z-scores), whereas no significant effect was observed on tongue movement for vowel /i(ː)/. See Figure 5.

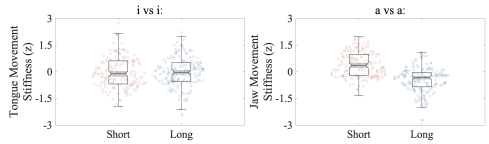


Figure 5: Stiffness of tongue movement for vowel /i(ː)/ (left) and of jaw movement for vowel /a(ː)/.

On the other hand, both tongue movement of vowel /i(ː)/ and jaw movement of vowel /a(ː)/ demonstrates a significant decreases when the utterance duration increases (for /i/: $t(326) = -6.15, p < 0.001, Est. = -0.45$ z-scores; for /a/: $t(333) = -5.69, p < 0.001, Est. = -0.32$ z-scores). We did not observe any significant effect of the interaction, indicating that the difference in stiffness between short and long vowels remains stable regardless of utterance duration.

3. Relationships of target and duration

After observing that long vowels exhibit a more extreme articulatory target compared to short vowels, characterized by higher maximum tongue height for vowel /i(ː)/ and lower minimum jaw height for vowel /a(ː)/, we proceeded to test the target undershoot model. 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.

3.1. Methods

To test this prediction, we extracted an additional measurement from the same dataset: duration to release (the duration from vowel articulatory onset to vowel articulatory release). 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 release, and their interaction. Subject was once again included as a random intercept in the analysis.

3.2. Results

Our findings reveal different behavior for the tongue height target of vowel /i(ː)/ and the jaw height target of vowel /a(ː)/.

For the tongue height target of vowel /i(ː)/, the effect of vowel length is significant ($t(323) = 5.16, p < 0.001$). The estimated effect size is approximately 0.60 z-scores, indicating that at the same average duration to release (0 z-score), long vowels exhibit higher maximum tongue height than their short counterparts by 0.60 z-scores. Furthermore, we did not observe any significant effect of duration to release or the interaction of vowel length and duration to release. This lack of significance indicates that the maximum tongue height and the difference in the tongue height target of short and long vowels remain stable across different duration to release. See Figure 6 (top).

On the contrary, for the jaw height target of vowel /a(ː)/, we did not observe any significant effect of vowel length, indicating that at the same average duration to release 0 z-score), long and short vowels do not have different minimum jaw heights. However, we found that the duration to release has a significant effect on the minimum jaw height ($t(330) = -5.61, p < 0.001$) with an effect size estimated at -0.71 z-scores. The negative effect size indicates that the minimum jaw height of short vowels decreases when the duration to release increases. Furthermore, the interaction between vowel length and duration to release is also significant ($t(330) = 2.18, p = 0.03$) with an effect size estimated at 0.32 z-scores. The positive effect size, combined with the larger negative effect size of duration to release, indicates that although the minimum jaw height of long vowels also decreases when the duration to release increases, the rate of de-

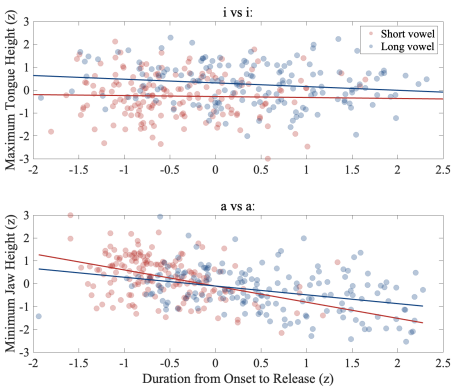


Figure 6: Tongue height for vowel /i(ː)/ (top) and jaw height for vowel /a(ː)/ (bottom) as a function of duration from onset to release of the vowel gesture. Solid lines represent regression line for each vowel length category.

crease for long vowels is not as high as their short counterparts. See Figure 6 (bottom).

4. Discussion and conclusion

Our articulatory investigation reveals not only durational differences but also distinct spatial features characterizing long and short vowels in Thai. Specifically, long vowels exhibit longer articulatory steady state duration, more prominent articulatory target, larger movement amplitude (only for jaw movement of vowel /aː/), and lower stiffness (only for jaw movement of vowel /aː/).

While our results, such as the observed differences in articulatory target and articulatory steady state duration, initially appear compatible with the target undershoot model, 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, at least for the jaw movement of vowel /aː/), seem to be category-specific, consistent with previous findings on singleton vs. geminate stops (Löfqvist 2005).

Furthermore, for the case of tongue height movement for vowel /i(ː)/, even when short and long vowels have an equal time to reach the target before the release, they still exhibit distinct articulatory target. Lacks of interaction between vowel length and duration to release also indicate the stability of target differences regardless of speech rate. These findings resonate with articulatory studies of vowel length in other languages, such as Australian English (Ratko, Proctor, and Cox 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.

One intriguing issue arises from the different behavior of tongue movement and jaw movement. Specifically, we observed a difference between short and long vowels in their articulatory targets when they have the same duration to release, which is only evident for tongue movement of vowel /i(ː)/, but absent for jaw movement of vowel /a(ː)/. We may interpret this as short /i/ and /iː/ having separate spatial targets, while short /a/ and long /aː/ share the same target, and the observed differences in spatial properties (as discussed in Section 2) are the result of undershoot.

However, the target undershoot model cannot fully explain the differences in dynamic kinematic properties, such as the disparity in peak velocity and stiffness, which are present for jaw movement of vowel /a(ː)/. The distinction between /i(ː)/ and /a(ː)/ may therefore stem from the fact that they involve different articulators, leading to differential behavior. Future studies are needed to explore the full set of articulators for vowels with different heights.

5. References

Abramson, Arthur S. (1962). *The vowels and tones of standard Thai: Acoustical measurements and experiments*. Indiana: Bloomington.

Abramson, Arthur S. and Nianqi Ren (1990). "Distinctive vowel length: Duration vs. spectrum in Thai". In: *Journal of Phonetics* 18.2, pp. 79–92.

Burroni, Francesco, Sireemas Maspong, Nicole Benker, Philip Hoole, and James Kirby (2024). "Spatiotemporal Features of Bilabial Geminate and Singleton Consonants in Italian". In: *Proceedings of the 13th International Seminar on Speech Production*.

Garrett, Andrew and Keith Johnson (2013). "Phonetic bias in sound change". In: *Origins of Sound Change: Approaches to Phonologization*. Oxford University Press. DOI: 10.1093/acprof:oso/9780199573745.003.0003.

Harrington, Jonathan, Philip Hoole, and Ulrich Reubold (2012). "A physiological analysis of high front, tense-lax vowel pairs in Standard Austrian and Standard German". In: *The Italian Journal of Linguistics* 24, pp. 149–173.

Hoole, Philip and Christine Mooshammer (2002). "Articulatory analysis of the German vowel system". In: *Silbenschnitt und Tonakzente*. Ed. by Peter Auer, Peter Gilles, and Helmut Spiekermann. Max Niemeyer Verlag, pp. 129–152. DOI: 10.1515/9783110916447.129.

Lindblom, B. (1963). "Spectrographic Study of Vowel Reduction". In: *The Journal of the Acoustical Society of America* 35.11, pp. 1773–1781. DOI: 10.1121/1.1918816.

Löfqvist, Anders (Jan. 2005). "Lip kinematics in long and short stop and fricative consonants". In: *The Journal of the Acoustical Society of America* 117.2, pp. 858–878. DOI: 10.1121/1.1840531.

Luangthongkam, Theraphan (2011). *Thai sounds: An acoustic study*. Bangkok: Chulalongkorn University Centenary Academic Development Project.

Mooshammer, Christine, Philip Hoole, and Anja Geumann (June 2007). "Jaw and Order". In: *Language and Speech* 50.2, pp. 145–176. DOI: 10.1177/00238309070500020101.

Onsuwan, Chutamanee (2005). "Temporal relations between consonants and vowels in Thai syllables". Ph.D. dissertation. University of Michigan.

Ratko, Louise, Michael Proctor, and Felicity Cox (Dec. 2023). "Articulation of vowel length contrasts in Australian English". en. In: *Journal of the International Phonetic Association* 53.3, pp. 774–803. DOI: 10.1017/S0025100322000068.

Roengpitya, Rungpat (2001). "A study of vowels, diphthongs, and tones in Thai". Thesis.

Tilsen, Sam and Mark Tiede (2023). "Parameters of unit-based measures of speech rate". In: *Speech Communication* 150, pp. 73–97. DOI: 10.1016/j.specom.2023.05.006.

Van Son, R. J. J. H. and Louis C. W. Pols (Oct. 1990). "Formant frequencies of Dutch vowels in a text, read at normal and fast rate". en. In: *The Journal of the Acoustical Society of America* 88.4, pp. 1683–1693. DOI: 10.1121/1.400243.

Variability in the articulation of Beijing Mandarin rhotic vowels

Song Jiang[1], Alexei Kochetov[1]

[1]*Department of Linguistics, University of Toronto, Canada*

soong.jiang@mail.utoronto.ca, al.kochetov@utoronto.ca

Abstract

One of the most documented characteristics of the North American English rhotic /ɹ/ (including its vocalic variant [ɚ]) is its contextual and/or inter-speaker variability in the choice of tongue shapes – bunched or retroflex. In contrast, the situation with Mandarin rhotic vowels (e.g. [ɚ, uˠ]) is much less clear. To further explore the individual and contextual variability, we have been conducting an extensive ultrasound investigation of Beijing Mandarin rhotic vowels. Preliminary results presented in this paper show both individual and contextual variation in the realization of these segments. First, the speakers we examined varied in using either a retroflex or a bunched tongue configuration. Second, we found some within-speaker variation conditioned by vocalic contexts, albeit not observed systematically. Third, the data also showed that rhotic vowels in Beijing Mandarin tend to be more similar to each other compared to their non-rhotic counterparts. These results demonstrate a greater than previously reported variability in the articulation of Beijing Mandarin rhotic vowels.

Keywords: rhotic vowels, Beijing Mandarin, ultrasound, individual variation, contextual variation

1. Introduction

One of the hallmark characteristics of the North American English rhotic /ɹ/ is its contextual and/or inter-speaker variability in the choice of tongue shapes. Delattre and Freeman's (1968) X-ray study illustrated eight general types of tongue shapes for the English /ɹ/, as produced by American English speakers. These tongue shape types can be grouped into two main allophonic lingual configurations – bunched and retroflex. Ong and Stone (1999) were first to report on vocalic contexts affecting /ɹ/'s lingual articulation based on ultrasound data: the rhotic tended to be bunched when flanked by front vowels. Mielke, Baker, and Archangeli (2016)' ultrasound study showed a high complexity of /ɹ/-allophony (bunched vs. retroflex) not only at the contextual level but also at the individual level. As /ɹ/-allophony appears to lack perceptible difference, speakers tend to adopt idiosyncratic articulatory strategies to reach the same acoustic goal (*i.e.*, F3 lowering for /ɹ/). However, variation at individual and/or contextual levels does not necessarily occur across languages. For example, Hussain and Mielke (2021) showed that rhotic vowels in Kalasha were found to be predominantly bunched regardless of their primary qualities (height and backness). Investigating rhotic sounds is therefore essential, as this provides insights into contextual variability and language- and speaker-specificity.

In contrast to English, the articulation of Mandarin rhotic vowels (underlying: /ɚ/ [ɚ] 'bait' or *er*-suffixed diminutive: /tu/-*er* [tuˠ] 'picture-DIM') is much less understood. Lee (2005) reported electromagnetic articulography (EMA) results showing exclusively tip-down (bunched) articulations for Beijing Mandarin (thereafter BM) rhotic vowels. Jiang, Chang, and Hsieh's (2019) EMA also illustrated that Northeast Mandarin speakers predominantly used a bunched lingual configuration in their productions of rhotic vowels. In contrast, Xing (2022) found predominantly tip-up (retroflex) tongue shapes through an ultrasound tongue imaging method. Chen and Mok's (2021) ultrasound study, however, found that Mandarin rhotic vowels can be articulated with either a retroflex or bunched configuration. Crucially, none of the studies have reported vowel-specific variability in tongue shapes, which is different from the results 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.

2. Methods

2.1. Participants

Six BM speakers (4 females, average age: 22.3; 2 males, average age: 21) were recruited for this experiment, which is part of a larger ongoing study. They were born in Beijing City and came to Canada after the age of 18.

2.2. Stimuli

The stimuli comprised of meaningful words with the vowels /u, ə, a/ and their *er*-suffixed counterparts [uˠ, ɚ, aɚ] preceded by bilabial stops, as listed in Table 1. [a] in BM is not rhotacizable, so an additional [ɚ] is realized in its diminutive.

Table 1: *Stimuli presented to speakers*

	Root	***er*-suffixed diminutive**
/u/	pu 'no'	puˠ 'step-DIM'
/ə/	pʰən 'gush'	pʰɚ 'basin-DIM'
/a/	pa 'to tyrannize'	paɚ 'handle-DIM'

2.3. Procedure

The experiments were conducted at the University of Toronto Phonetics Lab. The participants were asked to complete a demographic questionnaire and a pre-test screening. For the ultrasound task, they were asked to produce the target words in the carrier phrase "__, *mà* __ *ba*" ("__, curse with the word__") five times.

Ultrasound and audio data were collected using an EchoB system (Articulate Instruments Ltd.), set at a frame rate of 60 fps and a field of view of 103.2°. An UltraFit headset (Articulate Instruments Ltd.) was used to stabilize the probe during imaging. Audio-ultrasound synchronization was implemented in AAA software (Articulate Instruments Ltd.).

2.4. Analysis

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 *t*1-*t*7) and converted to polar coordinates.

To understand how the tongue moves and its shape changes over time, plots of temporal tongue contour changes were generated for each target rhyme using a custom Matlab script. The plots show tongue contours averaged over five repetitions within each speaker.

Polar Generalized Additive Mixed Models (thereafter GAMMs) were used to compare the tongue shapes between different rhymes within each speaker (following Heyne, Derrick, & Al-Tamimi 2019). GAMMs were fit to our polar data using *bam*() from *mgcv* package in R. In our GAMMs, we included rhymes as the fixed factor and tongue distances from the origin (radius) as the outcome variable. Angle values (radian) by token were included for factor smooth interaction. Predicted smooths were visualized using *plotly* package.

3. Results

3.1. Inter-speaker variation

Figure 1 shows temporal tongue contour changes in each speaker's production of [u˗], with the first and the last frame of the selected interval shown in dark blue and red, respectively. BM01, BM04, and BM05 used a retroflex configuration: the tongue tip was raised, while the tongue dorsum maintained an [u]-position. The other three participants, BM02, BM03, and BM06, used a bunched or 'tip-down' tongue shape to produce [u˗] with a domed-up tongue blade and a concave tongue back. While previous studies transcribed this sound as a monophthong (Shi 2009), all BM speakers in our dataset, except for BM06, showed a considerable change in the tongue shape over time.

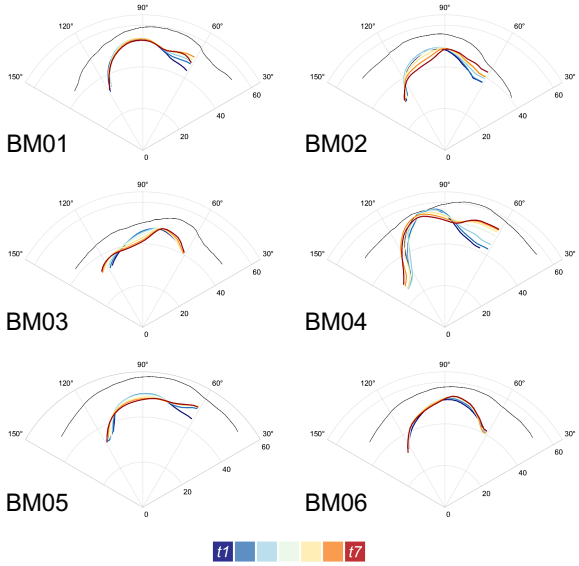


Figure 1: *Temporal tongue shape changes in the productions of [u˗].*

Figure 2 shows the temporal tongue contour changes in each speaker's production of [ɚ]. BM02 and BM03 used a prototypical bunched tongue shape to produce this sound: the tongue body was domed up, the tongue tip was held low, and a concavity was created in the dorsal region. BM05 used a front-bunched configuration (following Lawson *et al.* 2013's classification) with a raised tongue front and a concavity in the back. BM01, BM04, and BM06 used a retroflex tongue shape: the tongue tip was curled up during the course. BM01 and BM06 also lowered their tongue body, while BM04's tongue body remained relatively static. Similarly to [u˗], noticeable tongue shape changes can be seen for all six speakers' productions of [ɚ]. BM03 had a more static tongue shape throughout the vowel with a slight horizontal movement.

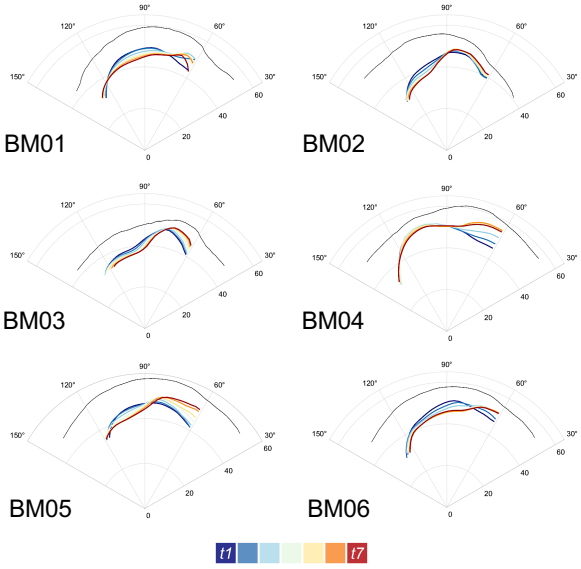


Figure 2: *Temporal tongue shape changes in the productions of [ɚ].*

Figure 3 shows temporal tongue shape changes for each speaker's production of the entire [aɚ] rhyme. BM02, BM03, BM05, and BM06 used a bunched tongue shape. The other two speakers, BM01 and BM04, used a retroflex configuration. In line with the transcription, our results illustrated a transition of the tongue shape from [a] to [ɚ], except for speaker BM06, who used different strategies in the productions of [ɚ] and [aɚ].

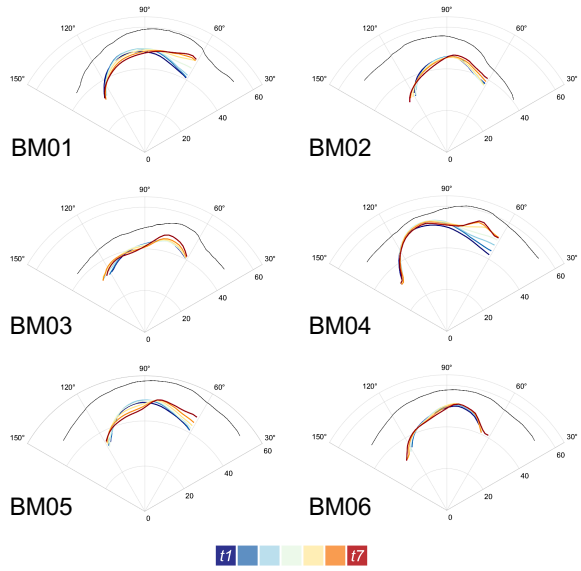


Figure 3: *Temporal tongue shape changes in the productions of [aɚ].*

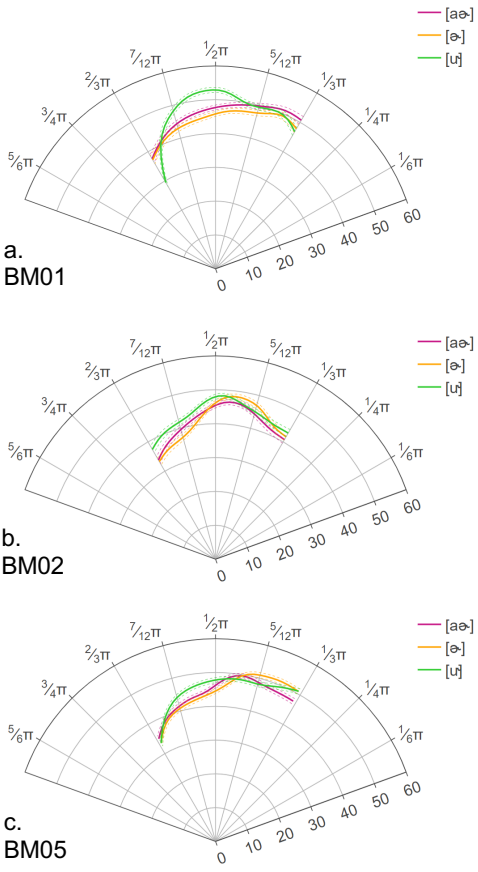
Overall, the results revealed considerable individual variation and, to some extent, within-speaker variation in the articulation of rhotic vowels. The speakers' lingual configurations are summarized in Table 2. The speakers varied in using either a retroflex or a bunched configuration. Four out of six speakers in our data (BM01, BM02, BM03, and BM04) used a single configuration for all three vowels, whereas BM05 and BM06 used both configurations for different vocalic contexts. In the following section, we will examine within-speaker variation further.

Table 2: *Summary of six speakers' configurations (R: retroflex; B: bunched)*

	BM01	BM02	BM03	BM04	BM05	BM06
uɻ	**R**	**B**	**B**	**R**	**R**	**B**
ɚ	**R**	**B**	**B**	**R**	**B**	**R**
aɚ	**R**	**B**	**B**	**R**	**B**	**B**

3.2. Within-speaker variation

To investigate the within-speaker contextual variation, tongue contours for [uɻ], [ɚ], and [aɚ] were compared using GAMM within each participant. The last frame *t*7 was selected to represent the time point of the maximum constriction based on the temporal data in Section 3.1. Figure 4a shows results from a representative speaker BM01 who used a retroflex lingual configuration across the board. The tongue tip was raised to similar positions for all three vowels. [uɻ] had a rather high tongue dorsum for BM01 in order to preserve the high-back vowel quality of [u]. Figure 4b shows the tongue shapes of speaker BM02 who used predominantly a bunched lingual configuration. It can be seen that all three vowels ended up with similar tongue shapes and positions. Despite the resemblance, [uɻ] had a slightly further back tongue position compared to [ɚ] and [aɚ]. [aɚ] had a lower tongue front and a more retracted tongue body than [ɚ], but the difference was subtle. As can be seen in Figure 4c, BM05 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. Similarly to BM02's articulation, [aɚ] had a more retracted tongue body compared to [ɚ]. BM06 in Figure 4d, 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. Moreover, BM06's [uɻ] had a higher tongue dorsum than [aɚ].



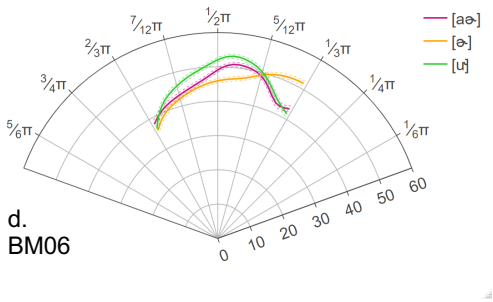


Figure 4: *GAMM plots of tongue contours for [aɚ], [ɚ], and [uɻ] at t7 (extremum point)*

To sum up, four of our six speakers showed within-speaker consistency in lingual configurations regardless of the vowel. Despite adopting the same configurations, these speakers still exhibited some context-conditioned articulatory variation. We found that [uɻ] tended to have a higher tongue dorsum than [ɚ]; a bunched [ɚ] tended to be retracted when preceded by [a]. The other two speakers showed within-speaker variation. The two speakers' vowel-specific strategies, however, were not the same, showing some idiosyncratic lingual configurations.

3.3. Rhotic vs. non-rhotic

Figure 5 and Figure 6 illustrate the GAMMs results for the mid-point (*t4*) of the non-rhotic vowels and the end-point (*t7*) of the rhotic ones from one retroflexing speaker and one bunching speaker. Results revealed that the tongue contours were much less spaced out for the rhotic vowels compared to their non-rhotic counterparts. For both speakers, either the tongue tip or the tongue blade was raised, while the tongue body was lowered, resulting in similar tongue positions. For the retroflex variants, the difference in dorsal height was preserved between [uɻ] and [ɚ], while diminished for the bunched variants. Only the backness contrast was observed in our dataset.

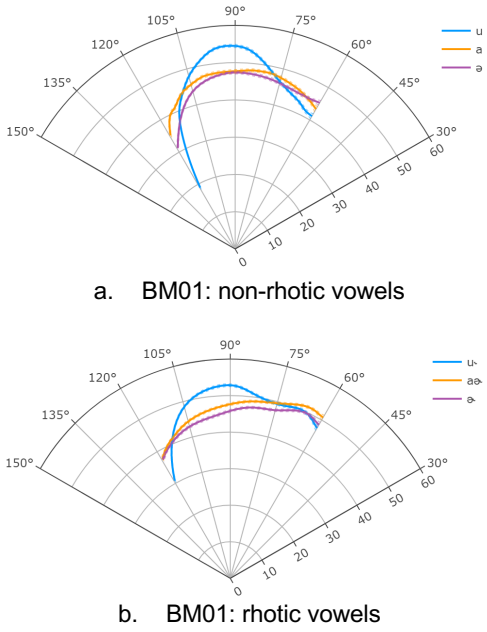
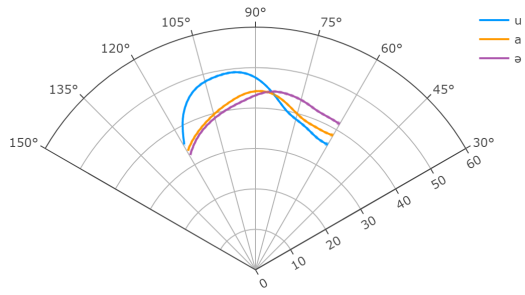
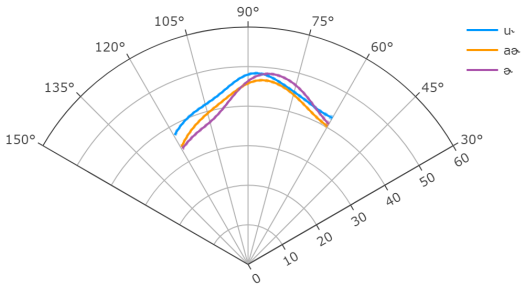


Figure 5: *GAMMs of the non-rhotic vowels [u, a, ə] (upper) and the corresponding rhotic forms (lower) produced by a retroflexing speaker – BM01*



a. BM02: non-rhotic vowels



b. BM02: rhotic vowels

Figure 6: *GAMMs of the non-rhotic vowels [u, a, ə] and the corresponding rhotic forms produced by a bunching speaker – BM02*

4. Discussion and conclusion

The 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, which echoes Chen and Mok's (2021) findings. Both configurations were about equally used by our participants. However, this contradicts the claim in the previous literature that rhotic vowels cross-linguistically prefer a bunched tongue shape (Mielke *et al.* 2016, Huang *et al.* 2024, among others). Huang *et al.* (2024) introduced the concept of 'gestural economy' (Maddieson 1995) to demonstrate that Southwest Mandarin rhotic vowel /ɚ/ is predominantly bunched because the vowel and consonant systems of this variety lack the retroflex gesture. BM, however, is well-known for its retroflex apical vowel and retroflex sibilants, which in the gestural sense bias BM speakers towards adopting retroflexion as a strategy for producing rhoticity. Xing's (2021) study suggested that retroflexion was predominantly employed in the articulation of BM rhotic vowels [u˞] and [ɤ˞]. This is different from the present study and may be due to Xing's choice to have retroflex fricative [ʂ] and affricate [tʂ] as onsets in the stimuli. These consonants could have led to the preference for retroflexion in the following rhotic vowels.

Second, we also found some within-speaker variation conditioned by vocalic contexts, although this was not observed systematically. This finding, 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 contextual variation. Our two speakers who showed contextual variation, however, did not adopt the same vowel-specific strategy. This shows that BM speakers can adopt idiosyncratic strategies in rhotic production, reminiscent of the /ɹ/-allophony in English (Mielke *et al.* 2016).

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 findings for 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. Our findings suggest that the complexity of BM rhotic vowels is manifested at both individual and contextual levels. Various factors can bias speakers' choice of a lingual configuration, such as consonantal environments or existing gestures in the inventory. BM rhotic vowels, nevertheless, show greater complexity in terms of the tongue postures than those in Kalasha and Southwest Mandarin as BM speakers tend to adopt more idiosyncratic strategies in rhotic production. This therefore highlights the need for a larger-scale investigation of rhotic sounds in BM (as well as across languages), with these sounds produced in a variety of phonetic contexts and lexical items. This work is currently underway.

5. Acknowledgements

This study has been approved by the University of Toronto Research Ethics Board #31791.

6. References

Chen, S., & Mok, P. P. K. (2021). Articulatory and acoustic features of Mandarin /ɹ/: A preliminary study. In 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP) (pp. 1-5). IEEE.

Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics, 6*(44), 29–68.

Heyne, M., Derrick, D. & Al-Tamimi, J. (2019). Native language influence on brass instrument performance: An application of generalized additive mixed models (GAMMs) to midsagittal ultrasound images of the tongue. *Frontiers in Psychology, 10*. 2597.

Huang, J., Hsieh, F., Chang, Y. & Tiede, M. (2024). On the two rhotic schwas in Southwestern Mandarin: when homophony meets morphology in articulation. *Phonetica, 81*(1), 43-80.

Hussain, Q., & Mielke, J. (2021). An acoustic and articulatory study of rhotic and rhotic-nasal vowels of Kalasha. *Journal of Phonetics, 87*, 101028.

Hussain, Q. & Mielke, J. (2022). The emergence of bunched vowels from retroflex approximants in endangered Dardic languages. *Linguistics Vanguard, 8*(s5), 597-610.

Jiang, S., Chang, Y. & Hsieh, F. (2019) An EMA study of er-suffixation in Northeastern Mandarin monophthongs. In Proceedings of 19th international congress of phonetic sciences, Melbourne, Australia 2019. Canberra: Australasian Speech Science and Technology Association Inc.

Lawson, E., Scobbie, J. M., & Stuart-Smith, J. (2013). Bunched /r/ promotes vowel merger to schwar: An ultrasound tongue imaging study of Scottish sociophonetic variation. *Journal of Phonetics, 41*(3-4), 198-210.

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.

Maddieson, I. (1995). Gestural economy. In Kjell Elenius & Peter Branderud (eds.), Proceedings of the 13th International congress of phonetic sciences, vol. 4, 574–577. Stockholm: KTH & Stockholm University.

Mielke, J., Baker, A., & Archangeli, D. (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language, 92*(1), 101-140.

Shi, F. (2003). Acoustic expression of er-ized finals in Beijing Mandarin, *Nankai Linguistics, 2*, pp. 11-19.

Xing, K. (2022). *Phonetic and phonological perspectives on rhoticity in Mandarin.* [Doctoral dissertation: The University of Manchester]

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

Kornélia Juhász, Huba Bartos

HUN-REN Hungarian Research Centre for Linguistics, Hungary
Eötvös Loránd University, Hungary

juhasz.kornelia@nytud.hun-ren.hu, bartos.huba@nytud.hun-ren.hu

Abstract

This preliminary analysis aims to shed light on how tonal coarticulation surfaces in the production of learners with an atonal L1, that is, Hungarian learners of Mandarin Chinese. In Mandarin assimilatory carry-over coarticulation is more salient, compared to dissimilatory anticipatory effect, besides, the realization of lexical tones and coarticulation also interacts with sentence type (i.e., intonation). Thus, in this analysis, we aim to investigate how the concatenation of lexical tones surfaces in two sentence types, i.e., declarative and syntactically marked yes/no interrogatives in the production of Hungarian learners of Mandarin and a native speaker group. All combination of the four Mandarin lexical tones appeared in the recorded tone sequences. F_0-contours of utterance-initial sequences of 4 syllables were compared by GAMMs. The results show that although L2 learners' f0 curves differ from the native realizations, but carry-over effect can be observed if the target tone is either T1 or T4 (i.e., the affected tonal target is H), whereas the same effect on T2 and T3 targets (with initial L tones) is not in line with native patterns, that is, does not show clear sign of trigger-dependent carry-over influence.

Keywords: Mandarin Chinese, tonal coarticulation, L2 production, atonal L2 learners' lexical tones

1. 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 1990a), yet show much weaker (dissimilatory) influence on the preceding lexical tone (Xu 1997). Based on the salience of carry-over effects, in this study we primarily focus on this progressive coarticulatory influence, yet our results include the analysis of anticipatory effects as well. If a carry-over effect is exerted between two adjacent lexical tones, then the interaction of the tonal targets is as follows: the low offset tonal target of the 1st lexical tone in the sequence – in an assimilatory manner – lowers the onset tonal target of the subsequent tone; likewise a high offset tonal target of the 1st tone elevates the subsequent tone's onset tonal target (Xu 1997). Likewise, in the case of anticipatory effects, the low onset of the 2nd tone regressively elevates the preceding tone's onset (in a dissimilatory manner), while the high onset in the same position lowers it. Additionally, local lexical tonal variations are not exclusively dependent on the neighboring tonal patterns, but also interact with sentence type (i.e. intonation). Thus, in this particular experiment, tonal coarticulation is observed in two sentence types, i.e., in declarative and syntactically marked yes-no interrogative intonation patterns. Sentence type itself primarily exerts influence on f0 register, however it might also alter f0 range, as well. According to Shen's MC intonation model (1990b), the declarative f0 curve displays a declining pattern, while yes-no interrogatives are marked by a significantly higher f0 throughout the whole utterance and may be characterized with a terminal rise (compared to the declarative contour). In the case of declaratives, anticipatory dissimilation is assumed to serve as counteract to declination in order to distinguish tonal patterns from the declination contour (Xu 1993: 122). In contrast, interrogatives are not characterized by a declining structure (Shen 1990b), hence different coarticulation-induced effects are expected here, relative to declaratives. Since the realization of lexical tones is altered not just as a result of tonal coarticulation, but also shaped by the interaction between lexical tones and sentence type, in this preliminary analysis we aim to shed light on how tonal coarticulation surfaces (if surfaces) in declarative and syntactically marked yes/no interrogatives in the production of Hungarian learners of Mandarin. Hungarian is an atonal L1, thus we hypothesize that concatenating lexical tones to sequences poses problems for Hungarian L2 learners, since the sequencing procedure requires the covariation of the above-mentioned factors, which are either absent or different from their L1. Additionally, concerning L1 intonation patterns, Hungarian declaratives are realized with a descending contour similar to MC, however, the prosodic structure of the character contour in yes/no questions is characterized by a rise followed by a fall (L*HL), which is initiated on the last accented syllable of the utterance (Varga 2002). This means that although MC and Hungarian interrogative patterns differ, but neither question curve is shaped by a gradual decline (in contrast to declaratives), which means that even if L1 transfer occurs in L2 production, the two sentence types is expected to induce different patterns of tonal coarticulation. In particular, in this analysis, we are seeking answers to the following questions: (i.) Does tonal coarticulation surface in atonal L2 learners' production? (ii.) And if so, does coarticulation interact with sentence types (i.e., declarative and syntactically marked yes/no interrogative sentence types)?

2. Methods

We analysed two adult speaker groups (5 female speakers per group): 1. Hungarians with cca. one year language experience (lower intermediate level) of MC: undergraduates majoring in Chinese ('L2 learners'); and 2. a control group of Chinese native speakers, who were born and raised near Beijing. All of the L2 learners use English (as a foreign language) on a daily basis. We recorded short question–answer dialogues, projected on a screen with both Chinese characters and pinyin transcriptions. Target sentences consisted of 4 (declaratives) or 5 words (interrogatives) (see Table 1). Question–answer pairs were recorded with 5 repetitions, in this manner we analysed 800 pairs in total (4 verbs × 4 objects × 5 repetitions × 10 speakers). The analysed tonal sequences were utterance-initial, consisting of 4 syllables in interrogative and declarative broad focus sentences, serving as SVO, and were followed by phonologically unspecified, weak syllable(s) (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. This means that carry-over tonal coarticulation was triggered by the 2nd syllable, affecting the 3rd syllable's target tonal realization (in which position all lexical tone appears). In contrast, anticipatory coarticulation surfaces in the opposite direction: the 3rd syllable regressively influences the 2nd syllable. Furthermore, coarticulatory effect appears between the 1st and the 2nd syllable, as well, but in this case the analysed tonal combinations are more limited. It must be noted, that the 4th syllable within the sequence bears a neutral tone, which also affects the lexical tone in the 3rd syllable position in a manner that if the full tone in the 3rd syllable is characterized by a dynamic tonal target (i.e. T2 (LH) or T4 (HL)), then the neutral tone in the 4th syllable takes the last target component of the corresponding preceding full tone (Shen 1992), but realized in a lower register, as opposed to full lexical tones, as a result of – left-to-right tonal spreading from the preceding full tone (Yip 1980). Consequently, the target approximation of the second component of the dynamic tones are expected to emerge in the 4th syllable of the sequence. Since the recorded utterances exclusively consisted of sonorants, f0 was 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 (Nolan 2003) in R (R Core Team 2019). F0 values were time-normalized syllable-wise (computing the relative position of the 5 ms steps within the duration of the syllable at hand), in order to be comparable by GAMMs (generalized additive mixed models) (Wood 2017) using the packages mgcv (Wood 2011) and itsadug (van Rij *et al.* 2022). In GAMMs the f0 change was analysed dependent on the normalized duration of concatenated syllables, besides, the model was complemented by an ordered parametric term (with contrast treatment) coding the speaker group and sentence type and merged into one single variable, reference curve set to native interrogative realization in each case. In total, we composed four GAMMs. Furthermore, random smooth function was applied in each case by to each f0-trajectory. The models were also treated for autocorrelation. Additionally, we carried out a qualitative analysis on the f0 curves in the following way: we determined the f0 of corresponding first tonal target of the 3rd syllable (i.e., the maximal or minimal excursion of the f0 curve associated with the lexical tone in the 3rd syllable) by 1 semitone intervals, and the f0 values within one interval were rounded downwards (e.g., an inflection point realized with an f0 value between 22 and 23 semitones is considered as 22 semitone). If there were no inflection points in the interval at hand, then the minimal (T2 & T3) or maximal (T1 & T4) f0 were considered.

Table 1: *The recorded interrogative and declarative utterances (the analysed sequences are in bold).*

Subject	**Verb**	**Object**		PRT	Q PRT
T1	T_y	T_x	N_1	N_2	N_3
他 *Tā* 'he'	**拉 *lā* 'pull'** **拦 *lán* 'hold back'** **理 *lǐ* 'understand'** **骂 *mà* 'scold'**	**妈妈 *māma*** 'mother' **爷爷 *yéye*** 'grandpa' **奶奶 *nǎinai*** 'grandma' **妹妹 *mèimei*** 'sister'		了 *le* (particle)	吗 *ma* (question particle, exclusively appearing in interrogatives)

3. Results

As regards to native Mandarin speakers' production interrogative f0 curves were in general elevated to a higher f0 register compared to declaratives, as expected, which in the majority of the cases meant significant discrimination between the two sentence types' f0 curves ($p < .001$). As for the declarative f0 curve, declination (in this case, the significantly lower f0 curve compared to the interrogative pattern) was initiated from the 2nd syllable of the sequence. Although the interrogative curve was elevated to a higher f0 level compared to the declarative pattern, the f0 curves' shape was almost identical in the two sentence types, in other words, the target approximation of lexical tones (i.e. the temporal alignment of inflection points within the two f0 curves) was alike in the two sentence types (Fig. 1). As regards to carry-over tonal coarticulation, the modification of the lexical tone in the 2nd syllable significantly influenced the realization of the fixed tone in the 3rd syllable (see each column on Figure 1., respectively). According to our qualitative comparison on carry-over effect in natives' production, exerted on the onset tonal target in the 3rd syllable (presented in Table 2), it is apparent that if the trigger tone's offset (i.e., the 2nd syllable's last tonal target element) is H (either T1 or T2), then the approximation of the next tonal target element is realized higher compared to those cases where the trigger tone's offset tonal target element is L (i.e., T3, T4). These patterns appear in native speakers' production irrespective of tonal sequence, as well as sentence type, which means that observed in both declaratives and interrogatives (Table 2). Additionally, the dissimilatory effect of anticipatory coarticulation is also observed is tendencies, e.g., the L target of T3 in the 3rd syllable elevates the preceding H in T1 (see the 1st row of Fig. 1). In contrast to native production, L2 learners' interrogative and declarative f0 curves were not differentiated significantly, rather the two f0 contours were overlapping in the majority of cases (Fig.1). Moreover, considering the f0 range of the tones individually in the sequence, it may be concluded that L2 learners produced more compressed tonal realizations (compared to native production), concatenated in a gradually descending pattern irrespectively to sentence type. 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 range, the discrimination of different tonal patterns positioned to the 2nd syllable was still present, approximating native patterns,

mainly in the case of T1 and T4. In contrast, approximating the native T2 and T3 patters – even in the 2nd syllable – posed problems to L2 learners. As regards to the carry-over effect expected to surface on the onset tonal target element of the 3rd syllable, the differentiation between the 4 lexical tones was challenging to L2 learners, which means that they struggled to produce tonal patterns as distinct as those of natives, adding up to overlapping f0 curves subsequent to different triggering contexts (Fig. 1). Concerning the results of the qualitative comparison of carry-over effects – in line with the

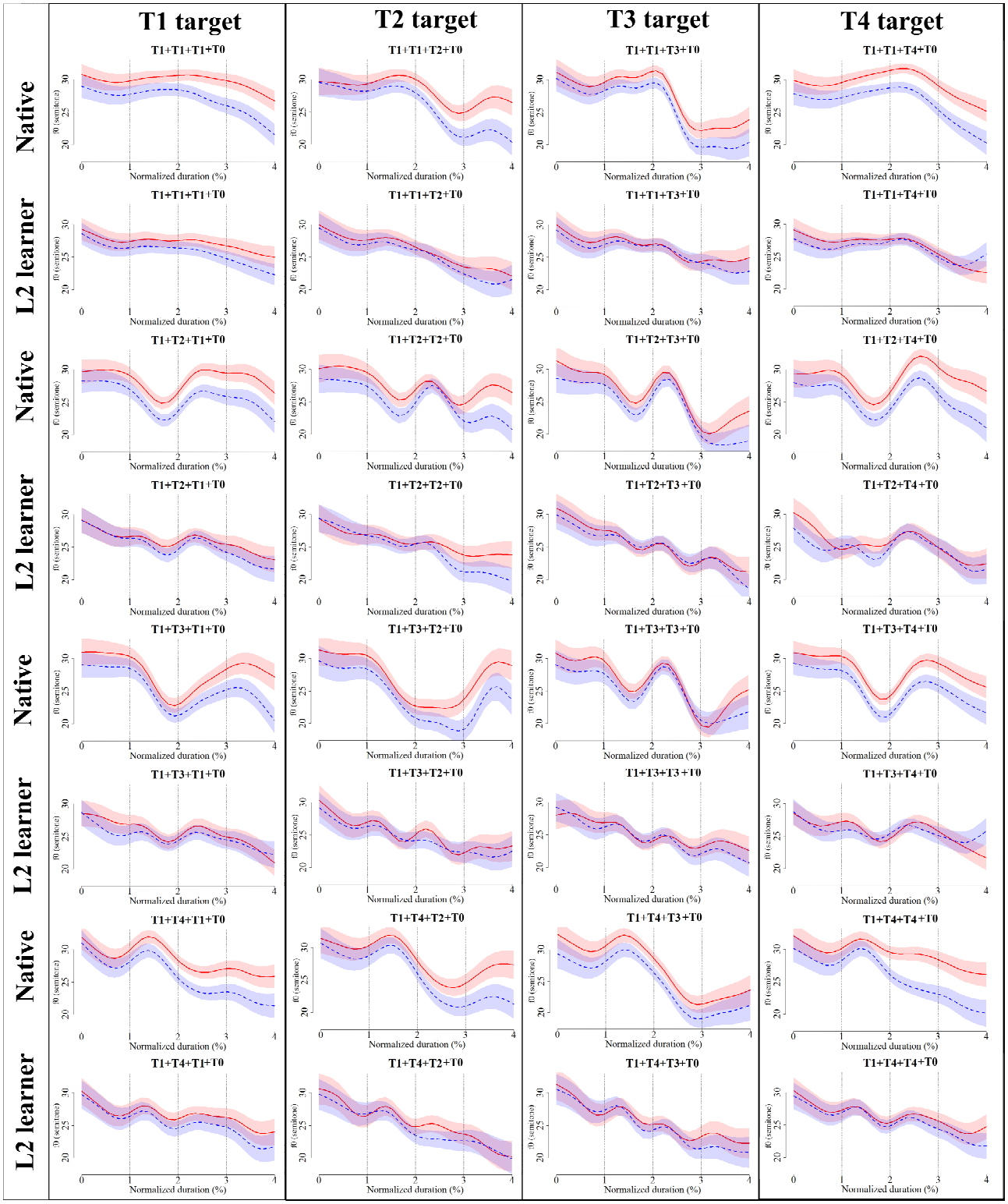


Figure 1: *The f0 curves of the different quadrisyllabic tonal sequences, where column-wise the 3rd syllable, row-wise the 2nd syllable is characterized by identical tonal value (i.e., the 2nd syllable triggers carry-over coarticulation on the target tone positioned to the 3rd syllable), and solid red line represents native, while dashed blue line represents L2 learners' estimated f0 pattern.*

observation presented above on the f0 curves – the data in Table 2 also confirms that if relative patterns are viewed exclusively between the realizations of lexical tones subsequent to different triggers, L2 learners more likely produce native-like patterns, if the subsequent target tone is either T1 or T4, irrespective of sentence type (Table 2). In particular, this means that if the tonal target element (which is influenced by the preceding trigger) is H, then the carry-over effect was similar to those of native speakers, (i.e., H trigger elevated, while L trigger lowered the subsequent target element's realization). In other cases, that is, if the L tonal target element is affected by the assimilatory carry-over effect, then the different tones realize almost identically, and there is no clear pattern to be observed based on the triggering contexts. However, it also must be noted that the T1 lexical tone in trigger position (i.e., in the 2nd syllable) is always induces relatively high f0, compared to all other three trigger tones in L2 learners production, similarly to native production. As regards to anticipatory effects in L2 learners' production, only tendencies and no clear patterns were observed.

Table 2: *Qualitative comparison of the carry-over effect in interrogative (I) and declarative (D) sentence types, where columns named as "Target Tx" show the one-semitone-interval of the f0 excursion associated with the first tonal target element of the lexical tone in the 3rd syllable, while "Trigger" denotes the lexical tone in the 2nd syllable exerting the effect. The interacting adjacent tonal target elements are shown in the column of "Interacting Targets".*

	Trig-ger	Interacting Targets offset-onset	I Target T1	D Target T1	Interacting Targets offset-onset	I Target T2	D Target T2
Native	T1	HH	30	27	HL	24	22
	T2	HH	29	26	HL	24	22
	T3	LH	28	25	LL	23	19
	T4	LH	28	23	LL	22	21
L2 learner	T1	HH	27	26	HL	23	22
	T2	HH	26	26	HL	23	21
	T3	LH	26	25	LL	22	22
	T4	LH	26	25	LL	23	22
		Interacting Targets offset-onset	I Target T3	D Target T3	Interacting Targets offset-onset	I Target T4	D Target T4
Native	T1	HL	23	19	HH	32	28
	T2	HL	20	18	HH	32	28
	T3	LL → HL	19	20	LH	29	26
	T4	LL	22	19	LH	29	23
L2 learner	T1	HL	24	24	HH	28	27
	T2	HL	22	23	HH	28	27
	T3	LL → HL	23	22	LH	27	26
	T4	LL	23	22	LH	27	26

4. Discussion and conclusion

In this preliminary analysis we aimed to shed light on how tonal coarticulation surfaces in the production of Hungarian learners of Mandarin. Since tonal realizations are highly dependent on intonation as well, we observed two sentence types (i.e., declarative and syntactically marked yes/no interrogative patterns) in the production of a L2 learner group, compared to Chinese natives. We analysed the initial quadrisyllabic interval of the tonal sequence, in which all combination of the four Mandarin lexical tone appeared. In the analysed sequences, the 2nd syllable was the trigger of which offset tonal target element was expected to influence the subsequent tonal target element (i.e., the onset target) of the lexical tone positioned in the 3rd syllable in an assimilatory manner, since our primary focus was placed on the more salient carry-over effect, in contrast to the less significant dissimilatory anticipatory influence. The obtained f0 curves were compared by GAMMs. Our results show, that L2 learners failed to produce the native-like discrimination of lexical tones in the two sentence types, since L2 learners' interrogative and declarative f0 curves were characterized by overlapping patterns (as opposed to native production). Furthermore the distance between maximal and minimal f0 of individual tonal realizations was also compressed relative to native patterns. Our result show that the carry-over effect at hand was affected L2 learners' production in a native-like manner, however only if the target tone (receiving the carry-over effect) had a H onset tonal target element (i.e., T1 or T4). In the case of T2 and T3 (where the onset tonal target element is L) the native-like coarticulatory patterns did not surface, rather different triggers induced almost identical effects on the onset tonal target element of the 3rd syllable). These results suggest that lexical tones with H onset targets are easier to produce and concatenate in a native-like manner. One explanation to this phenomena might be rooted in physiological reasons: the low f0 register is more difficult to reach, as it requires more articulatory efforts. (For example, in Mandarin Chinese, owing to these constraints low f0 register shows less flexibility as well, as regards to the realization of L targets, most apparently in an utterance-final position (Xu 1993). Consequently, in general, it might be assumed that L2 learners produce tonal interaction effects less sophisticated and clear in the low f0 register due to the articulatory limitations of voicing. Besides, it must be added that L2 learners' production is characterized by positive and negative excursions on the f0 curve, however these excursions did not approximate the natives' magnitude. This could mean that the compressed realizations and transitions between targets could limit the processes counteracting declination (both in the case of declaratives, as well as interrogatives), resulting in limited f0 change in the lower f0 register. These preliminary results shed light on problems of lexical tone production and tone sequencing in the case of Hungarian learners of Chinese. The results could also contribute for further investigations focusing on tonal coarticulation in the production of atonal learners of Mandarin.

5. Acknowledgment

This paper reports initial results of the research carried out in the collaborative project under grant no. NKM2023-15 of the Hungarian Academy of Sciences and the Chinese Academy of Social Sciences: 'The acoustic analysis of Mandarin Chinese tones and intonation in the production of Hungarian learners of Chinese.'

6. 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 F_0 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.

Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. *Proc. of 15th ICPhS,* 771-774.

R Core Team, *R: A language and environment* f*or statistical computing.* R Foundation for Statistical Computing. v. 3.6.1. Available: https://www.R-project.org/. 2019.

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

Shen, X. (1990b) *The Prosody of Mandarin Chinese*. California: University of California Press.

Shen, X. (1992). Mandarin neutral tone revisited, *Acta Linguistica Hafniensia,* 24, 131-151.

van Rij, J., Wieling, F., Baayen, R. H., & van Rijn, H. (2022). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.* R package version 2.4.1.

Varga, L. The intonation of monosyllabic Hungarian yes-no questions, *Acta Linguistica Hungarica*, 49(3-4), 307-320.

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

Wood. S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models", *Journal of the Royal Statistical Society,* 73(1), 3-36.

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.

Xu. Y. (1993). Contextual tonal variation in Mandarin Chinese. Ph.D. Dissertation. The University of Connecticut.

Yip, M. (1980). The tonal phonology of Chinese. PhD dissertation. MIT.

Effects of phonetic contexts on aerodynamic conditions for uvular trills in French

Andrés Felipe Lara[1], Didier Demolin[2], Claire Pillot-Loiseau[3]

[1]*Laboratoire de Phonétique et Phonologie (CNRS et U. Sorbonne Nouvelle), Paris, France*

[andres.lara, didier.demolin, claire.pillot]@sorbonne-nouvelle.fr

Abstract

This study examines the impact of vowel quality and syllabic position on the aerodynamic requirements for producing the French uvular trill. The findings suggest that using the vowel [a] when producing rhotics is more likely to result in trills, whether in word-initial or intervocalic positions. This is followed by [u], while rhotics produced with [i] tend to favor fricatives rather than trills. Trill production is more advantageous in word-initial positions compared to intervocalic positions. Furthermore, the results demonstrate that as the duration of sustained productions above a 2 hPa threshold increases, the conditions for trilling become more favorable in the [a] and [u] contexts. The recommended time limits for trilling in these contexts are 110 ms for [a] and 140 ms for [u]. Additionally, voiceless trills are significantly longer than other modes of articulation for French rhotics (i.e. approximants and fricatives), suggesting that duration can be employed as a distinguishing factor for voiceless trills.

Keywords: aerodynamics, coarticulation, rhotics

1. 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 thresholds reaching 3.2 hPa as observed by 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).

2. 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 (see figure 1). Beats observed were then manually counted after the initial rise and dip of intraoral pressure (Po).

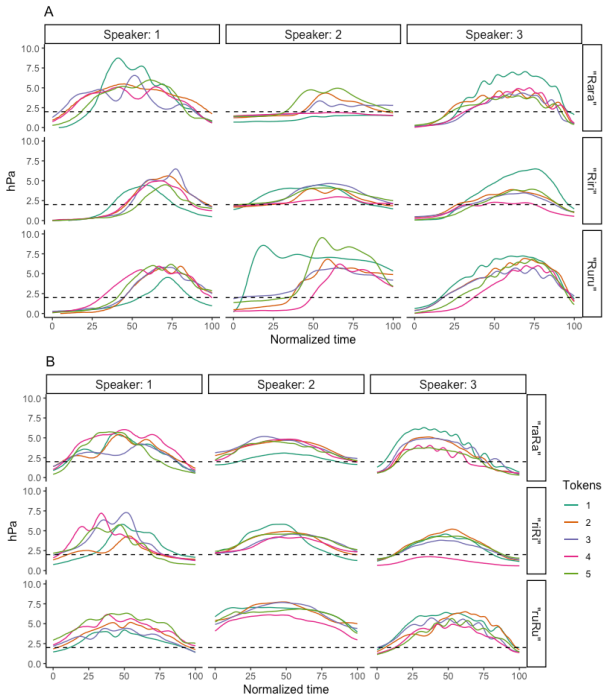


Figure 1: *Lines represent intraoral pressure measured in hPa (hectopascals). Measurements for three speakers, with speaker 3 identified as female. Each data point is further categorized by word position (A: word-initial, B: intervocalic) and token (represented by color). A dotted line marks the 2hPa threshold.*

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 et al., 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). Mode classification was also employed in the search for indications in the

identification of trills from a purely acoustic perspective. Given that the signal underwent filtration, solely the parameters of duration and fundamental frequency (F0) were extracted as a means to facilitate the aerodynamic-acoustic comparison. A two-way analysis of variance (ANOVA) was conducted to investigate significant differences between trills, fricatives, and approximants in terms of duration and voicing. Following a statistically significant ANOVA, a Fisher's Least Significant Difference (LSD) test was performed at a 99% family-wise confidence level to determine differences between means.

3. Results

See figure 2 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.

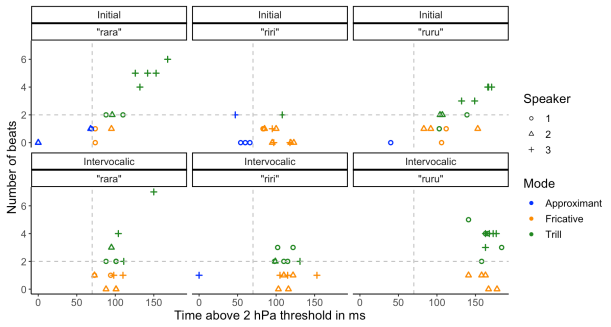


Figure 2: *Rhotic productions in initial and intervocalic contexts (all tokens included). Productions are categorized by production mode (colors) and speaker (shapes). x-axis: duration (in ms) the Po is maintained above the 2 hPa threshold, with a reference line at 70ms. y-axis: number of oscillations achieved.*

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 3 demonstrates the posterior probabilities for the model.

Figure 3 shows predictions from the Bayesian model fitted for categorical regression between the three modes of articulation. 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.

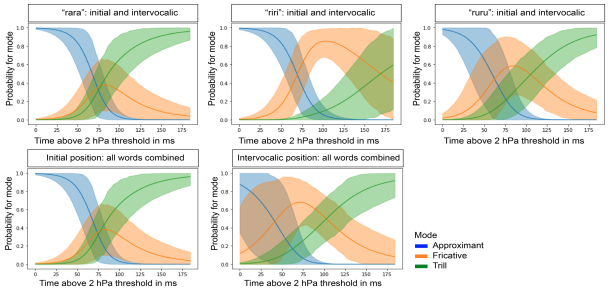


Figure 3: *Categorical regression showing posteriors for the mode of production of rhotics focusing on vocalic context (top tier) and the distinction between initial and intervocalic positions (lower tier). y-axis: probability of each mode;1 indicates highest probability relative to Po being sustained above 2 hPa: x-axis.*

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.

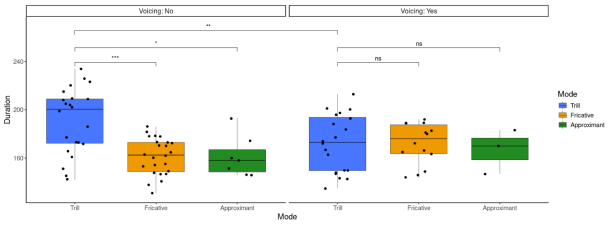


Figure 4: *Differences in duration and voicing among various renditions of the rhotic are examined. Significance levels (* < .05, ** < .01, * < .001, ns = no significance) are presented to compare the trill mode with two other modes, as well as to compare a voiced trill with a voiceless one.*

Figure 4 shows comparisons between the different modes of articulation based on voicing and duration. No significant findings were observed when comparing the modes of articulation in voiced productions. The duration was found to be approximately equal for all modes of articulation in voiced productions. However, there was substantial variation (in terms of duration) identified in trills across voiced and voiceless productions. In voiceless productions, a highly significant comparison ($F(5, 84) = 2.583$, $p = .00057$) was found between trills and fricatives. Additionally, a significant effect on duration ($F(5, 84) = 2.583$, $p = .019$) was observed when comparing voiceless trills and devoiced approximants. Trills were found to be significantly longer than both fricatives and approximants. Furthermore, an examination of the differences between trills in voiced and voiceless productions revealed highly significant differences ($F(5, 84) = 2.583$, $p = .0018$).

Notably, voiceless trills were significantly longer than voiced trills. The devoiced approximants shown in Figure 4 were not classified as fricatives because of their low Po values. In general, devoiced approximants did not exceed 2.65 hPa.

4. Discussion and conclusion

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]. Statistical analyses revealed significant differences in duration between voiceless and voiced trills, with voiceless trills presenting longer durations. This study provides support for the findings of Solé (2002), which suggest that voiceless trills are more robust compared to voiced trills, thus making voiceless trills easier to sustain. Longer durations, which indicate the ability of a speaker to sustain a trill, are associated with voiceless trills and exhibit a higher number of beats (Lewis, 2004; Solé, 2002). Furthermore, voiceless trills were found to be significantly longer than other modes of articulation, indicating that duration can serve as a distinguishing factor for trills when produced without voicing. This research is of great significance as it sheds light on the phonetic and phonological complexities of trills, providing valuable insights for the categorization and analysis of rhotic sounds in French. It also contributes to the development of articulatory modeling and synthesis.

5. Acknowledgements

This work is part of the program "Investissements d'Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL).

6. 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.

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@gipsa-lab.grenoble-inp.fr

Abstract

Somatosensory and auditory feedback contribute to speech motor control, but it is unclear how they interact in on-line feedback control. In previous studies, we showed evidence for a somatosensory-based response to tongue-stretch perturbation in vowel production, which ensures tongue posture stabilization and preserves the auditory characteristics of the sound. In this study, we combined the tongue perturbation with an alteration of the auditory feedback, which induced formant shifts that were either consistent or inconsistent with the auditory impact of the tongue perturbation. We investigated how the compensation for the auditory perturbation interacts with the somatosensory response to the tongue perturbation. We did not find any interaction. The latency of the compensation for the formant shift was longer than the one of the somatosensory responses, suggesting that somatosensory feedback control could be the fastest one to preserve crucial auditory characteristics of vowels.

Keywords*: speech motor control, on-line feedback mechanism, mechanical perturbation, reflex*

Introduction

Speech is auditory in nature. Hence, auditory feedback is crucial to achieve speech goals and precise speech production (Savariaux et al., 1999; Perkell et al., 2000; Purcell & Munhall, 2006; Cai et al., 2011). However, somatosensory feedback has also been shown to play an important role both for speech motor control, (Tremblay et al, 2003, Nasir & Ostry, 2008) and for vowel identification in the absence of auditory feedback (Patri et al., 2020).

In a recent study (Ito et al 2020), using a sudden tongue-stretch perturbation during steady-state vowel production, we have found clear evidence for a quick on-line compensatory response (with a 130-ms latency) aiming at preserving the production of the vowel against the perturbation. We have also shown (Bourhis et al., submitted) that this compensatory response occurs similarly when the participants receive their normal auditory feedback and when their auditory feedback is masked by a pink noise. This result suggests a crucial role of somatosensory feedback in the generation of the observed response to the tongue-stretch perturbation. This was confirmed under the same experimental conditions by an EMG study of the muscles acting on the anterior part of the tongue: an increase of muscle activity was observed around 60ms after perturbation onset. This is a relatively short latency which is more compatible with polysynaptic somatosensory reflex, than with typical phonetic auditory correction (Ito et al., 2024).

Importantly, we observed that the compensatory response did not bring the tongue back to its position before the perturbation onset, but to another position that preserved the tongue contour in the constriction of the vocal tract and was compatible with the achievement of the crucial auditory characteristics of the vowel. This suggests that in speech production somatosensory feedback could be specifically tuned, so as to ensure accurate acoustic vowel production, even in the absence of auditory monitoring.

However, our results do not discard a possible role of auditory feedback, when it is available. Indeed, the condition of our experiment may not allow to demonstrate this contribution, since the somatosensory correction and the auditory correction act in the same direction, aiming at recovering the auditory characteristics of the produced vowel. Previous studies from the literature, using on-line alterations of the auditory feedback, both at the levels of the formants (Purcell & Munhall, 2006) and of the pitch (Larson et al, 2000) during steady-state vowel production, have shown latencies of the auditory correction that were longer (>200ms) than the latency of the response observed in our study (130ms). However, latencies of auditory corrections as short as 120ms were found when the perturbation was applied during the production of a sequence of vowels, i.e. under dynamical speech production conditions (Cai et al 2011, Xu et al, 2004, Donath et al 2002). Since our tongue-stretch perturbation induces a displacement of the tongue during vowel production, we cannot discard the possibility that an auditory correction mechanism associated with dynamical speech production, could also be involved in our steady-state production task.

In the condition of our study using tongue-stretch perturbation, in order for us to be able to detect the specific auditory contribution to the response of the tongue-stretch perturbation, it is necessary to break the compatibility between the somatosensory-based and the auditory-based corrections. This can be done by examining whether an auditory correction induced by altered auditory feedback is not similar to the somatosensory-based correction.

To address this, we carried out a somatosensory-auditory perturbation test by combining the tongue-stretch perturbation with an auditory perturbation. The auditory perturbation was applied to the first formant, and was either in the same direction as the acoustic consequence of the tongue-stretch perturbation (decrease of F1) or in the opposite direction (increase of F1). Based on the latencies found in the literature in steady-state vowel production, we predicted that the latency of the auditory-based correction could be longer than the one of the somatosensory corrections. Hence, the additional auditory error induced by the F1-shifts could not affect the quick compensatory response to the tongue-stretch perturbation. To verify the latency of auditory correction, we also tested auditory-perturbation alone conditions, which was applied to the first formant in both directions.

Method

Twelve native French speakers participated in the experiment. They reported no known speech or hearing impairment and no history of profound injury that could induce a somatosensory loss in the orofacial region. This experiment was approved by the local ethical committee (CERGA: Comité d'éthique pour la recherche, Grenoble-Alpes [CERGA-AvisConsultatif-2021-18]). All participants signed the consent form.

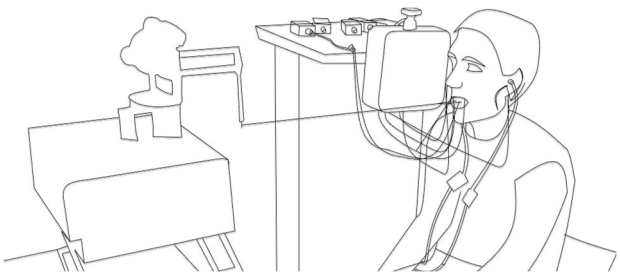


Figure 1: *Experimental setup*

The experimental setup is presented in Figure 1.

For the sensory perturbations, a tongue-stretch perturbation (PTB) and an altered auditory feedback perturbation (AAF) were used. For the tongue-stretch perturbation, we applied the same method as in our previous studies (Ito et al, 2020, Bourhis et al, submitted). A small robotic device (Phantom Premium 1.0, Geomagic) was connected to the tongue surface through a thin thread glued on both lateral sides of the tongue blade. A 1N force was applied in the forward direction as a step function with rise and fall phases of 5 ms, which prevents mechanical noise in the robot. For the auditory feedback perturbation, F1 was modified by 20 % either upward (incongruent with the effect of the mechanical perturbation) or downward (congruent with the effect of the mechanical perturbation) using Audapter (Cai et al., 2011). The altered sound was played back with 70 dB of white noise through magnetic compatible earphones (Natus Tip 300).

We recorded displacements of the tongue and jaw using electromagnetic articulography (Wave, Northern Digital Inc.). Six sensors were attached to the upper lip, lower lip, jaw, tongue tip, blade and dorsum in the mid-sagittal plane of the head. Reference sensors were also attached to the nasion, left and right mastoids, and the upper incisor for head movement correction. For each participant, the palate contour in the midsagittal plane was recorded by tracing the surface of the palate with a sensor glued on the experimenter's finger. The data were sampled at 200Hz. The produced speech sounds were also recorded using Audapter (Cai et al, 2011) at a 11.025kHz sampling rate: the first four formants, F1, F2, F3, and F4, were extracted at a sample frequency of around 345Hz.

In the test, the participants were asked to sustain vowel /ɛ/ for 3s in response to a visual cue. Vowel production started and ended with closed mouth position. Each trial was triggered manually by the experimenter after checking that the participant was ready. The two perturbations (PTB and AAF) were applied 1s after the onset of the vocalization. The tongue perturbation lasted for 1s. The auditory perturbation lasted until the end of the trial for a total duration of 4s. We tested five perturbed conditions combining auditory and tongue perturbations: 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 a pseudo randomly selected one third of the trials, so that the mechanical perturbation was never applied in two consecutive trials. Each of the five perturbed conditions was applied once within blocks of 15 trials. In total, 15 responses per condition were recorded (15 blocks).

We focused on the analysis of the acoustical data. Trials with wrong formant estimation (F1<300 Hz or F1>700 Hz) were removed from the analysis. Two participants were removed from the analysis due to high trial-to-trial variability. Acoustic data were aligned by the onset of the tongue perturbation, and they were averaged across perturbed trials in each condition and in each participant. To remove individual variability of F1 amplitude, F1 was normalized by dividing it with the baseline amplitude that is the value averaged over the 50ms interval preceding the onset of the auditory perturbation.

We first compared AAFup+PTB and AAFdown+PTB and assessed whether auditory feedback changes the compensatory response to the tongue-stretch perturbation. As shown in our previous studies (Ito et al. 2020, Bourhis et al. submitted), the tongue stretch perturbation induces a decrease of F1 and the compensatory response reduces this decrease. We compared the peak amplitudes of the initial decrease and of the time course of the compensatory response. The times of these peaks were obtained based on the average response calculated over the three conditions involving tongue-stretch perturbation (PTB, AAFup+PTB and AAFdown+PTB). The time points of interest are labeled as P1 and P2 in Figure 2. The peak amplitudes were calculated over 20ms windows centered at these time points.

To characterize the role of the auditory feedback, we also compared AAFup and AAFdown conditions and assessed whether these two responses diverged or remained similar over the course of the vowel production. For this analysis, we focused on two time points, namely T_{base}: at the baseline and T_{diff}: at the onset time of the divergence between the two formant responses (Figure 2). T_{base}: was set 150ms before the perturbation onset. To detect T_{diff} we applied a cluster-based analysis (Groppe et al, 2011). The procedure is based on a permutation test repeated 1000 times that was applied at each sampling point. We took in consideration the onset time of the first interval in which reliable difference was found over a set of consecutive sampling points (i.e a cluster). The amplitude at each time point was obtained using a 50ms window centered at this point.

A repeated measure ANOVA was applied in each amplitude comparison.

Results

We first compared the F1 responses to the tongue-stretch perturbation in two auditory conditions (AAFup+PTB and AAFdown+PTB). The normalized F1 responses in these two conditions are represented in the bottom panel of Figure 2. As in our previous studies (Ito et al, 2020, Bourhis et al, submitted), the tongue-stretch perturbation changed F1 and induced compensatory responses. The normalized F1 value decreased to about 0.83 116ms after the perturbation onset and the compensatory response brought it back to about 0.9 240ms after the perturbation onset. The temporal pattern of the responses is similar in two auditory conditions. Figure 3 represents the difference between the two auditory conditions in the normalized formant values measured at the times of the peak formant decrease (P1) and of the peak of the compensatory response (P2). We applied a two-way ANOVA on these formant values (time: P1 *vs* P2 and auditory condition: upshift *vs* downshift). There was no significant difference between auditory conditions ($p > 0.86$), but a significant difference exists in the time factor ($p < 0.001$). The interaction between the two factors was not significant ($p > 0.83$). These results indicate that a compensatory response was systematically induced by the tongue stretch-perturbation and that this response was not significantly affected by the auditory perturbation.

To verify the effect of AAF perturbation alone, we also compared the two auditory conditions without tongue-stretch perturbation (AAFup and AAFdown). The top panel of Figure 2 represents the normalized F1 response in these two

conditions. These two responses are similar in the time interval from the perturbation onset to time P2, and start diverging around 300ms. A cluster analysis reveals that this divergence starts from 360ms. Based on this analysis, we focused on the two time points T_{base} and T_{diff} shown in Figure 2.

The differences in normalized F1 value at these time points between the two auditory conditions in the absence of tongue stretch perturbation are represented in the left panel of Figure 3. A significant difference exists at T_{diff}, but not at T_{base}, indicating that F1 produced by the participants was significantly modified in response to its alteration induced by the auditory perturbation. As expected, this change was induced in a direction opposite to that of the perceived formant shift. The results also indicates that the compensation in response to the auditory perturbation was induced with a longer latency than the compensatory response to the tongue-stretch perturbation.

A divergence was also observed in the normalized F1 responses observed in the two auditory conditions in presence of the tongue-stretch perturbation (bottom panel of Figure 2). However, the difference at time T_{diff}, was not significant (see the right panel in Figure 4). The cluster based analyis also showed a difference occurring at a much later time (>1.2 s). This late difference detection may be due to the particularly large inter-participants variability associated with the tongue-stretch perturbation.

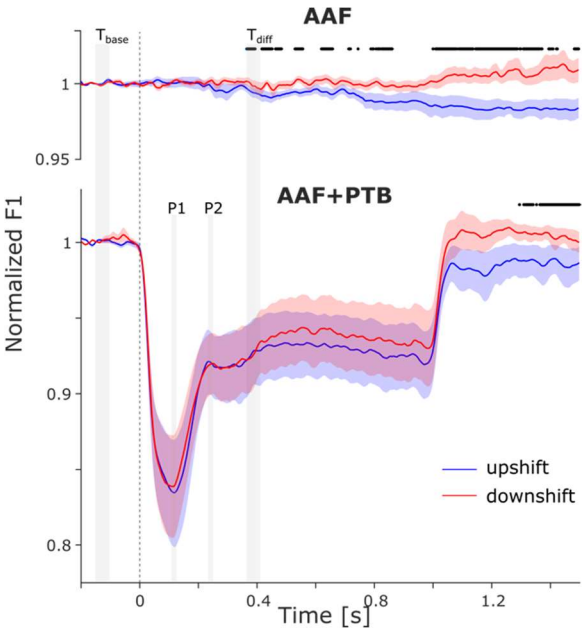


Figure 2: *Normalized F1 responses in AAF (top panel) and AAF+PTB (bottom panel) conditions. The colored shaded areas represent the standard errors across participants. The vertical grey bars represent the times for which a comparison between auditory conditions was made. The black dots at the top of each panel represent the sample points at which a significant difference between the auditory conditions was revealed by the cluster-based analysis. See methods for details.*

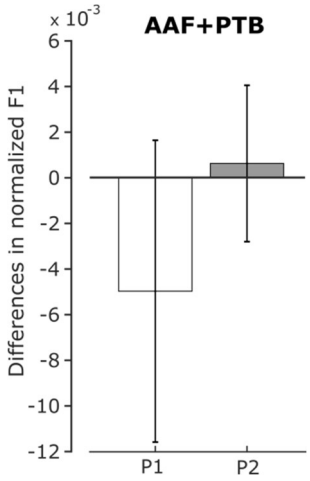


Figure 3: *Differences in normalized F1 values between the two auditory conditions in presence of the tongue stretch perturbation (upshift: AAFup+PTB and downshift: AAFdown+PTB) at focused time points of interest (P1 and P2 in Figure 2). P1 corresponds the peak of the initial decrease of F1 and P2 corresponds the peak of the compensatory response. The error bars represent the standard error across participants. See methods for details.*

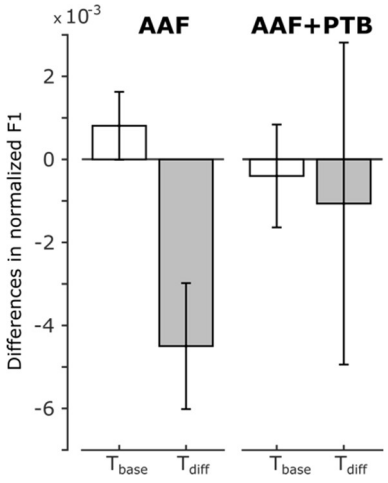


Figure 4: *Differences in normalized F1 value between the two auditory conditions (upshift and downshift) in the (left panel) and presence (right panel)à of the tongue stretch perturbation at two time points (T_{base} and T_{diff} in Figure 2). The error bars represent the standard error across participants. See methods for details.*

Discussion and conclusion

In this study, we combined a sudden tongue-stretch perturbation with a shift of the first formant F1 during the steady-state production of vowel /ɛ/. The formant shift was either in the same direction as the acoustic impact of the tongue perturbation or in the opposite direction.

The latency of auditory compensations for alterations of the auditory feedback involving formant shifts was shown (Cai et al., 2011) to be similar to the latency of the response to the tongue-stretch perturbation observed in our experiment, when

the auditory perturbation was applied during the production of a time-varying sequence of sounds (dynamical speech production henceforth). The decrease of the latency of auditory corrections in dynamical speech production, compared to static speech production, has also been observed in studies using pitch perturbation. The response latency was shorter when the pitch perturbation was applied during disyllabic sequences (100-150ms) (Donath et al., 2002; Xu et al., 2004) than the ones (150-200ms) when the perturbation was applied in the sustained vowel (Larson et al., 2000). Thus, the shortest latency of auditory-based corrections is comparable with the observed latency of the response to the tongue-stretch perturbation (around 130ms). Hence, in our experiment the auditory-based correction of the formant shift may influence the compensation for the tongue perturbation.

In line with previous studies (Ito et al 2020, Purcell et al, 2006), in our study both perturbations induced a compensatory response, and they were not simultaneous. The tongue-stretch perturbation induced a quick compensatory response with an average latency of 116ms. In contrast, the latency of auditory correction of the formant shift occurred significantly later, with an average latency of 360ms. Importantly the quick compensatory response to the tongue-stretch perturbation was not influenced by the additional formant shifts. These results confirm that the quick compensatory response relies on somatosensory feedback alone, and they suggest that somatosensory and auditory feedback control mechanisms may work separately and sequentially, due to their clearly different latencies.

Hence, our results show that, despite the tongue movement induced by the stretch perturbation during steady-state vowel production, the latency of the correction of the formant shift is the same as in usual steady-state vowel production (~ 400ms in Purcell et al., 2006). This suggests that auditory feedback is dependent on the planed speech production task (steady-state in our experiment) and not on whether or tongue movement occurs. We could expect different results if the tongue perturbation was applied during dynamical speech production. Just like the auditory-based corrections, the somatosensory response to perturbation could feature a significantly shorter latency.

In addition to the latency, the amplitudes of compensation were different between the two perturbations. The produced sounds were changed by about 15 % due to the tongue perturbation (see Figure 2) and 20 % due to the auditory perturbation. While the somatosensory-based compensation induced a recovery of around 50 % of the change induced by the tongue perturbation, the auditory compensation recovered only a few percent of the formant shift. Since in the tongue perturbation auditory change is associated with a compatible somatosensory error, it is easy to compensate for both sensory errors simultaneously. This is different from the case that the formant shift that is not associated with any somatosensory error.

Although the amplitude of compensation due to AAF perturbation was relatively small, this results still indicate the involvement of auditory error-detection mechanism. This detected error may be used in adaptation mechanism.

Overall, our results showed that somatosensory feedback induced faster compensatory responses than auditory feedback. For on-line speech motor control, these two compensatory mechanisms could be involved in different temporal phases, independently.

Acknowledgements

This work was supported by grants from the Agence Nationale de la Recherche (ANR-21-CE28-0022, PI. Takayuki Ito) and the National Institute on Deafness and Other Communication Disorders (Grant R01-DC017439).

References

Bourhis, M., Perrier, P., Savariaux, C., Ito, T. (Submitted). Quick speech motor correction in the absence of auditory feedback.

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.

Donath, T. M., Natke, U., & Kalveram, K. Th. (2002). Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. *The Journal of the Acoustical Society of America*, 111(1), 357–366. https://doi.org/10.1121/1.1424870.

Feng, Y., Gracco, V. L., & Max, L. (2011). Integration of auditory and somatosensory error signals in the neural control of speech movements. Journal of neurophysiology, 106(2), 667–679. https://doi.org/10.1152/jn.00638.2010.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event‐related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, 48(12), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x.

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.

Ito T, Bouguerra M, Bourhis M, Perrier P (2024) Tongue reflex for speech posture control. *Scientific Reports*, 14(1):6386.

Larson CR, Burnett TA, Kiran S, Hain TC (2000) Effects of pitch-shift velocity on voice F0 responses. J Acoust Soc Am 107:559–564.

Nasir, S. M., & Ostry, D. J. (2008). Speech motor learning in profoundly deaf adults. Nature Neuroscience, 11(10), 1217–1222. https://doi.org/10.1038/nn.2193.

Patri, J. F., Ostry, D. J., Diard, J., Schwartz, J. L., Trudeau-Fisette, P., Savariaux, C., & Perrier, P. (2020). Speakers are able to categorize vowels based on tongue somatosensation. *Proceedings of the National Academy of Sciences*, *117*(11), 6255-6263.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., ... & Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics*, *28*(3), 233-272.

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.

Savariaux, C., Perrier, P., Orliaguet, J. P., & Schwartz, J. L. (1999). Compensation strategies for the perturbation of French [u] using a lip tube. II. Perceptual analysis. *The Journal of the Acoustical Society of America*, *106*(1), 381-393.

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

Xu, Y., Larson, C. R., Bauer, J. J., & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. The Journal of the Acoustical Society of America, 116(2), 1168–1178. https://doi.org/10.1121/1.1763952.

Some Effects of Frame Rate on Gesture Detection in Tongue Ultrasound

Pertti Palo, Steven M. Lulich

Indiana University Bloomington

pertti.palo@taurlin.org, slulich@indiana.edu

Abstract

We study how decreasing ultrasound frame rate affects automated speech gesture detection. The gesture detection is performed on Pixel Difference contours with simulated stepping down of the frame rate from ≈122Hz down to ≈17Hz. We report how this affects the number of peaks detected and the accuracy of peak locations for Pixel Difference calculated using six different vector norms. The results point to a steady degradation of detection results as the frame rate is decreased.

Keywords: speech timing, speech gestures, sampling frequency, automated methods

1. Introduction

While tongue ultrasound is widely used in speech research and related areas, the analysis is often limited to selecting 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 (Wrench and Balch-Tomes 2022), and in the case of using holistic image based methods, we are no longer limited to basing the analysis on comparing single sample points. Instead, we can analyse articulation as a (almost) continuous function of time (Palo 2019; Al-Tamimi and Palo 2023). In time domain analysis, the sampling frequency or frame rate of the data becomes an important factor that can limit the analysis we are able to perform (Palo and Lulich 2023).

Palo and Lulich (2023) used a method called Pixel Difference (PD) for speech gesture analysis. PD evaluates the overall change in an ultrasound image sequence by interpreting the images as vectors and calculating the distance between consecutive images as a vector norm (Palo 2019). Similar methods have been used by, for example, Drake, Schaeffler, and Corley (2013) and Raeesy, Baghai-Ravary, and Coleman (2011). **Figure 1** demonstrates PD and the effect of lower frame rates on this type of analysis.

To state this problem broadly, we are interested in what the limit frequency is for speech articulation gestures to be detectable in articulatory data. More specifically, we will concentrate on tongue ultrasounds. The simple answer to this type of question in signal processing comes from the Nyquist-Shannon sampling theorem and states that to detect a signal without aliasing artefacts we need a sampling frequency that is at least double the frequency of the signal (Shannon 1949). However, we are going to need to do better than just detect a signal at the frequency of interest.

In a related study, Wrench and Scobbie (2008) used data from two different ultrasound systems. They sampled extracted tongue contours along two directions from the ultrasound probe origin to produce graphs of contour movement in the root and tip regions of the tongue. They show that 60 Hz data (produced by de-interlacing 30 Hz video ultrasound data) produces analysis results that are on par with those produced by a 100 Hz system showing a clear-to-the-eye difference between the /ele/-gesture in "Pay Laver" vs. "Pale Eva" in both types of data. On the other hand, spectral analysis of X-ray microbeam point tracking data shows that most of its information is in the under 12 Hz band as shown by correlation analysis with acoustic data (Goldstein 2019).

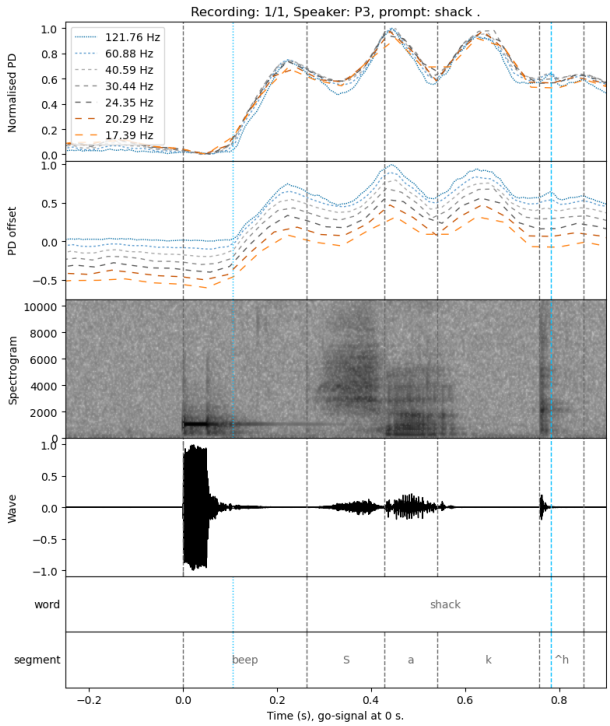


Figure 1: *Effects of downsampling on PD. Top panel shows PD curves resulting from no downsampling (121.76 Hz) up to a factor of 7 (17.39 Hz). The second panel shows the same data offset for clarity of individual curves. Bottom panels are the spectrogram, waveform (the go-signal beep is the part with the largest amplitude) and word with phonological segmentation. The vertical dotted blue line marks movement onset on the original data, and the vertical dashed blue line marks a gesture peak associated with aspiration of /k/.*

Analysing gesture timing without anatomical reference points is quite different from analysing contour movement and despite some commonalities - analysing flesh-point tracking like X-ray microbeam is different from PD analysis of image sequence data like ultrasound. For detailed analysis of speech timing from time series, we are going to need to be able to iden-

tify minima and maxima at a good enough accuracy and to not lose any that are produced by the gestures. In **Figure 1** we can see why the latter are a concern. Looking at articulatory onset (marked with vertical dotted blue line), we can see that at lower sampling frequencies we would not be able to identify it as accurately, which would quickly produce problems in study designs that need high statistical power. Furthermore, the gesture peak marked with a vertical dashed blue line disappears completely as sampling frequency goes down.

Continuing our recent work (Palo and Lulich 2023), we seek to empirically determine what the sampling frequency of tongue ultrasound needs to be in order for automatic peak detection to be able to find a believable number of gesture peaks in an utterance. To do so, we explore the effect of two variables on peak detection: the sampling frequency and the vector norm (a type of l_n^p-norm) used to calculate PD.

2. Materials

The data is a sample of 174 single-word utterances of a delayed naming experiment. The words were single-syllable lexical English words with a word final plosive ([p, t, k]) and an onset consonant ranging from none to /CCC/. The data was recorded at 121.76 fps in the mid-sagittal plane synchronised with audio. For details, please see Experiment 2, Participant 3 in Palo (2019). This speaker's data has good tongue surface visibility and provides a good baseline for this proof-of-concept study.

3. Methods

3.1. Downsampling

The original data is downsampled by a factor ranging from 2 to 7. For a downsampling factor of n this is done by using only every n^th frame in the ultrasound data for analysis. For example, for a factor of 3, we use frames [1, 4, 7, 10, ...] as the analysed data. Since the original data was recorded at 121.76 fps, this gives the sampling frequencies shown in **Figure 1**: 60.88, 40.59, 30.44, 24.35, 20.29, and 17.39 Hz.

3.2. Vector norms

Vector lp norms – or more precisely l_n^p-norms – can be defined as shown in Equation 1. In our case p is the order of the norm, n length of the vector or size of the ultrasound frame in pixels, and x_i are the individual elements of the vector, which in PD are evaluated as differences between corresponding pixels in consecutive frames.

$$l_n^p = \begin{cases} \sum_{i=1}^{n} \frac{|x_i|}{1+|x_i|}, & p = 0 \\ \sum_{i=1}^{n} |x_i|^p, & 0 < p < 1 \\ \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}, & 1 \leq p < \infty \\ \max(|x_i|), & p = \infty \end{cases} \quad (1)$$

Since the parameter n is defined by the number of pixels in the analysed frames, we will use the simpler notation of lp in the rest of this paper. 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.

3.3. Peak detection

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 process and produce reasonable accuracy in identifying actual gesture peaks. The process was guided by observing the results on a test set of 10 recordings for norms $l0.5$ $l1$, $l2$, and $l5$. The recordings were the first 10 in the data set.

A conservative lower limit for the gesture interval (parameter `distance`) was estimated from the data of Jacewicz, Fox, and Wei (2010). They report a high limit of approximately 6.7 syllables/second for speech rate (see Figure 1 in Jacewicz, Fox, and Wei (2010)). Given that syllables can be expected to have at least two gestures associated with them, we arrive at 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 by scaling it accordingly and rounding up.

The `width` parameter was chosen as 1 (meaning a peak with a width of 1 sample halfway down its prominence value was accepted as valid). The test set would have merited using a higher value if we were only interested in getting the best results for that set. However, using a higher value would make peak detection deteriorate very fast with downsampling as time spanned by 3 frames expands. We are still going to see degradation of the results when the sampling frequency gets close to the Nyquist frequency. This is actually desirable because the articulatory gestures are not sinusoidal signals, and in order to analyse them we need better time resolution than that required by the Nyquist frequency condition.

Finally, `prominence` was selected by stepping its value within the set (0.005, 0.01, 0.02, 0.03, 0.04). The last value was found to exclude peaks in the test set that we did not want to exclude, and so we used the value 0.03 for the `prominence`. It should be noted that while the individual parameters behave occasionally in an unintuitive manner, on the whole the way that `scipy.signal.find_peaks` works provides a very intuitive and easy to use way of identifying the peaks we are interested in.

3.4. Choosing the Period of Interest

Lower limit of the Period of Interest (POI) was set at 58 ms from the beginning of the go-signal (50 ms long 1 kHz sinusoidal beep) after Palo (2019) based on the minimal reaction times calculated by Chiu and Gick (2014). The upper limit of the POI was set the length of a gesture interval after end of the word to include the possible plosive release gesture in the analysis.

We used the chosen inter-gesture interval to extend the POI from the end of the utterance-final burst segment's *beginning* to account for cases where the plosive produced only a short release burst, and thus the acoustic boundary is at times already before the release gesture peak or very close to it.

4. Results

Our results are illustrated in **Figures 2-4**. Downsampling causes the number of detected peaks (**Figure 2**) to mainly decline for all of the norms with $l1$, $l2$, and $l5$ showing the best stability. However, the sample-by-sample peak number ratio distributions show that in some cases downsampling first increases the number of detected peaks as evident in that some distribution tails in **Figure 3** are above 1. This effect is strongest in $l5$ and $l\infty$.

Figure 4 shows that the peak position errors increase for

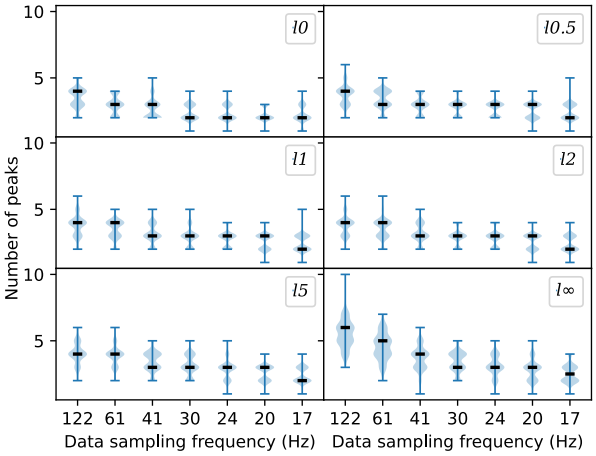


Figure 2: *Distributions of number of peaks detected in each sample. Black bars mark distribution medians.*

all of the norms while $l1$ and $l2$ behave the best in this respect. In this figure we relate the position errors to the limit set above in Section 3.3 for the minimum time between gestures: $t = 0.075$ s. All of the error distribution tails cross the limit already at 41 fps. At 30 fps and below there are more than outliers above the limit for each norm. None of the distribution medians cross the limit before 17 fps.

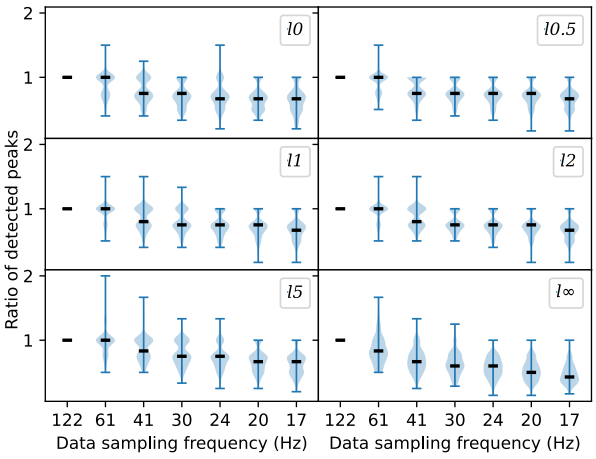


Figure 3: *Distributions of ratio of peaks detected in each sample, compared to those in the original data (122 Hz). Black bars mark distribution medians.*

5. Discussion

A necessary caveat on our results is that the approach we have taken is not exactly the same as using ultrasound with a lower frame rate. This is because the analysis here achieves a lower frame rate by dropping frames. As such it remains unclear if a longer frame acquisition time will affect the quality as well. This seems likely as longer frame acquisition means that the likelihood of within-frame movement artefacts increases.

As for the speech materials analysed, the data comes from only a single speaker. Further, it should be noted that our selection of phonetic content is limited. We did not have any flaps, taps, or trills in the test dataset. Of these flaps and taps are likely to be the fastest tongue gestures and should be included in a more comprehensive analysis. Trills on the other hand are held longer in terms of the whole tongue, and while they have a dynamic target, attaining the target can be imaged with a lower frequency than two times the trill frequency.

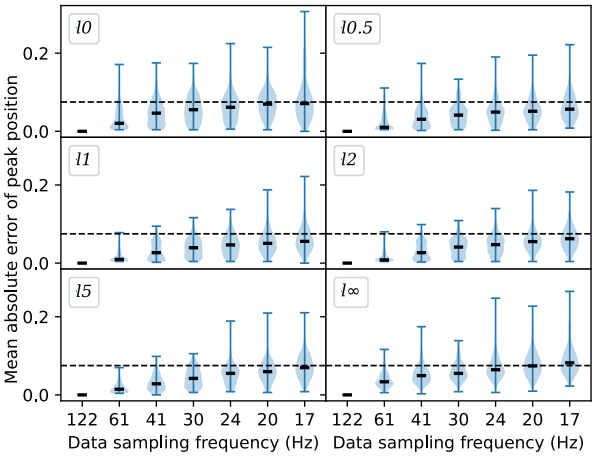


Figure 4: *Distributions of time accuracy of peak detection compared to the original data. Black bars mark distribution medians and the dashed line marks the lower limit of the time between gestures t=0.075 s used in the peak detection.*

6. Conclusion

The results show that quality of automatic peak detection degrades steadily with dropping of the sampling frequency. There does not seem to be any kind of division into two regions where the results would be good down to a given sampling frequency and then sharply change after that. Instead, the results point to a conclusion that a higher sampling frequency is always desirable.

There is no clear winner in terms of the used norms either. There is clear indication, however, that the limit norms – $l0$ and $l\infty$ should not be used. Rather, if there is an optimal norm or norm region, it will probably be somewhere close to $l1$ and $l2$.

As for the method itself, the results do show that useful, actionable information can be gained by this type of analysis. In particular, this study provides a reason to prefer high frame rates when using automated gesture detection. Before analysing a larger and more varied data set, the conclusion about lower frame rates of data must remain only a tentative caution.

7. Acknowledgements

Pertti Palo's work has been funded by a post-doctoral grant from the Emil Aaltonen foundation via the Post-Doc Pool of Finland.

8. References

Chiu, C and B. Gick (2014). "Startling Speech: Eliciting Prepared Speech Using Startling Auditory Stimulus". In: *Frontiers in Psychology* 5.1082.

Drake, E., S. Schaeffler, and M. Corley (2013). "ARTICULATORY EVIDENCE FOR THE INVOLVEMENT OF THE SPEECH PRODUCTION SYSTEM IN THE GENERATION OF PREDICTIONS DURING COMPREHENSION". In: *Architectures and Mechanisms for Language Processing (AMLaP)*. Marseille.

Goldstein, Louis (2019). "The Role of Temporal Modulation in Sensorimotor Interaction". In: *Frontiers in Psychology* 10. DOI: 10.3389/fpsyg.2019.02608. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.02608/full (visited on 04/16/2024).

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.

Palo, P. (2019). "Measuring Pre-Speech Articulation". PhD thesis. Edinburgh: Queen Margaret University.

Palo, P. and S. M. Lulich (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.

Raeesy, Z., L. Baghai-Ravary, and J. Coleman (2011). "Parametrising Degree of Articulator Movement from Dynamic MRI Data". In: *12th Interspeech*, pp. 2853–2856.

Shannon, C.E. (1949). "Communication in the Presence of Noise". In: *Proceedings of the Institute of Radio Engineers* 37.1, pp. 10–21. DOI: 10.1109/JRPROC.1949.232969. URL: https://ieeexplore.ieee.org/document/1697831 (visited on 04/09/2024).

Al-Tamimi, J. and P. Palo (2023). "Dynamics of the Tongue Contour in the Production of Guttural Consonants in Levantine Arabic". In: *International Conference of Phonetic Sciences (ICPhS 2023)*. Prague.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors (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.

Wrench, A. and J. M. Scobbie (2008). "High-Speed Cineloop Ultrasound vs. Video Ultrasound Tongue Imaging: Comparison of Front and Back Lingual Gesture Location and Relative Timing." In: *Proceedings of ISSP 2008 - 8th International Seminar on Speech Production*.

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, New Haven, CT, USA*

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

Abstract

Dynamical models of articulatory gestures relate the velocity of a vocal tract variable to its position via a function with one or more control parameters. In this paper we propose a minimal dynamical model of gestures. The model is empirically motivated by observations of the timecourse of the ratio of velocity to position in bilabial constriction movements by English and Mandarin speakers. We discovered that this ratio tends to follow an exponential growth curve over the course of a movement. A dynamical formalization of this empirical discovery, in combination with an assumption of point attractor dynamics, constitutes the core of our model. The model has only two parameters, T *and* r. T *corresponds to the target position of the vocal tract variable and* r *corresponds to rapidity. Simulations from the model capture key elements of gesture kinematics, performing much better than the damped mass-spring model. Our model achieves these improvements despite having fewer control parameters. Future work will extend our model to other kinds of gestures besides bilabial consonant constrictions.*

Keywords: *articulatory gesture, articulatory kinematics, dynamical system, damped mass-spring*

1. 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 distance to the target, of the form in (1).

$$\dot{x} = -\lambda(x - T) \tag{1}$$

x is the state of a vocal tract variable (TV) like lip aperture (LA: the distance between the lips), T is the target state of the TV (e.g., zero or possibly negative for /b/ or /m/ [Parrell, 2011]), and λ is a control parameter modulating the relationship between velocity $\dot{x}$ and distance to the target $(x - T)$. We follow Mücke et al. (2024) in using T instead of x_0 to refer to the target position, since x_0 often refers to the initial state of x. (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, peak velocity occurs later, approximately halfway through the movement (Ostry et al., 1987). In the *damped mass-spring* model of Task Dynamics, as in (2), peak velocity is delayed because velocity $\dot{x}$ is negatively related to acceleration $\ddot{x}$.

$$b\dot{x} = -k(x - T) - m\ddot{x} \tag{2}$$

The timing of the velocity peak predicted by (2) is an improvement over (1). This improvement is achieved via greater model complexity: (2) is a *second order* system, referencing acceleration in addition to velocity, with four control parameters m, b, k, and T, more than the two parameters λ and T in (1). Even in (2), however, peak velocity occurs unrealistically early (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 distance to the target (Sorensen & Gafos, 2016).

In this paper, we take a strongly empirical approach to understanding the relation between velocity and position. Rather than commit to the specific second order system in (2), we start from the minimal assumption that velocity is negatively related to distance to the target, formalized in (1). This allows us to solve for the parameter λ from measurement of data, in particular, electromagnetic articulatography (EMA) recordings of bilabial constriction movements. In this way, we address the question: what is the *empirical* relationship between velocity and position over time? The answer to this question guides further dynamical model development, which we pursue below.

2. Methods

2.1. Participants

Data was collected from 24 subjects: 12 native speakers of American English (8 female, 4 male, ages 19–28, mean = 20.75) and 12 native speakers of Mandarin Chinese (7 female, 4 male, 1 nonbinary, ages 19–33, mean = 24.00). All participants self-reported no history of speech, language, or hearing impairment.

2.2. Stimuli

Stimuli consisted of eight word-initial CV sequences in each language, where the initial consonant was bilabial—either [b] or [m]—and the vowel was either low back [ɑ] or high front [i]. Target sequences containing the vowel [i] were immediately preceded by the vowel [ɑ], and sequences containing the vowel [ɑ] were immediately preceded by the vowel [i], in order to ensure maximal vowel movement. All Mandarin target syllables bore a falling tone (T4) and were preceded immediately by a low tone (T3). Each target syllable was produced in two carrier sentences, occurring once in an informationally prominent position and once in a less prominent position. To encourage natural speech, each carrier sentence was preceded by a question, which served to provide context for the target sentences.

2.3. Procedure

Presentation of materials was controlled using E-Prime. On each trial, an audio recording of a question was played. The question was also displayed in text on the screen for 5000 ms.

Participants were instructed to listen to the question and to read aloud the answer that followed. In total, each participant produced 128 tokens (8 items × 2 carrier sentences × 8 repetitions) across four blocks of 32 items each. Within each block, stimuli were presented in a randomized order.

Articulatory kinematic data was collected with the NDI Wave Speech Research System sampling at a rate of 100 Hz. The sensors of interest for this study were attached at the vermillion border of the upper lip (UL) and lower lip (LL). Three sensors were also attached to the tongue: tongue tip (TT), tongue blade (TB), and tongue dorsum (TD), placed ~1 cm, ~3 cm, and ~5 cm from the tip of the tongue, respectively. In order to track movements of the jaw, one lower incisor (LI) sensor was attached to the hard tissue of the gum directly below the left incisor. Reference sensors were attached on the left and right mastoids and on the nasion. Measurements of the occlusal plane and a midsagittal palate trace were also collected. Acoustic data was collected using a Sennheiser shotgun microphone at a sampling rate of 22,050 Hz.

2.4. Data processing

Articulatory data was rotated to the occlusal plane and corrected for head movement computationally. Trajectories were smoothed using the robust smoothing algorithm of Garcia (2010). First and second time derivatives (velocity and acceleration) were calculated from the smoothed trajectory using central differencing, then lowpass filtered using a 5th order Butterworth filter. Consonant constriction gestures were parsed from the lip aperture (LA) signal, calculated as the Euclidean distance between the UL and LL sensors. The onset and offset of each movement were marked as the timepoints at which velocity exceeded or fell below, respectively, a 20% threshold of peak velocity, manually selected in MVIEW (Tiede, 2005). The spatial target of each gesture (i.e., T) was defined as the LA value at the timepoint of minimum velocity following gesture offset.

λ was calculated at each sample as the negative ratio of instantaneous velocity to instantaneous distance to the target: $-\dot{x}/(x-T)$ (see [1]). By demarcating gestures based on a 20% threshold of peak velocity, instead of, e.g., velocity zero-crossing, we exclude portions of the kinematics in which velocity or distance to the target are infinitesimal. This prevents λ from approaching 0 (infinitesimal velocity) or infinity (infinitesimal distance to the target). Gesture duration was calculated by subtracting the timestamp of the onset of movement from the timestamp of the offset of movement. We also calculated a measure of kinematic stiffness for each gesture by dividing peak velocity by maximum spatial displacement, i.e., onset position minus target position (Roon et al., 2021).

Out of the 3,072 tokens elicited, a total of 962 tokens (31.3%) were eliminated from analysis for the following reasons: failure of the gesture parsing tool to extract the gesture (447 tokens); a non-monotonic trajectory, i.e., instantaneous velocity changed sign for at least one sample (306 tokens); failure of the participant to produce contrastive focus on the informationally prominent syllable, as judged by the experimenters (155 tokens); disfluency (5 tokens); or data storage failure (49 tokens).

3. Results

3.1. Kinematic variables

Figure 1 displays the distributions of the kinematic variables gesture duration, peak velocity, maximum displacement, and kinematic stiffness across all 2,110 tokens from all 24 speakers.

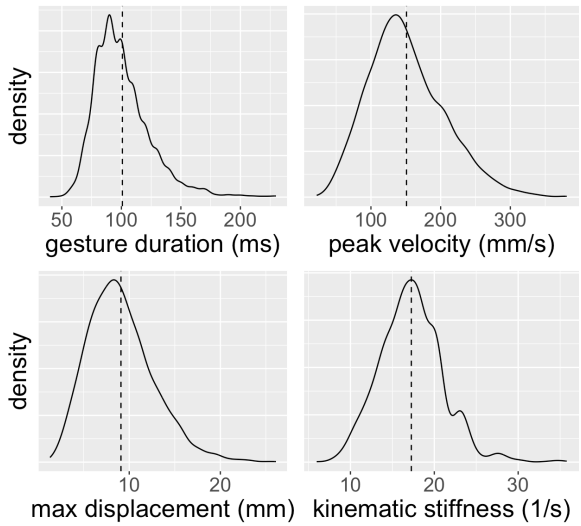


Figure 1: *Density plots of kinematic variables across all tokens (*n *= 2,110). Dashed vertical lines indicate the mean.*

3.2. λ trajectories

Next, we examine the trajectories of λ, i.e., the ratio of instantaneous velocity to instantaneous distance to the target. As seen in **Figure 2**, regardless of language and vowel context, λ generally followed an exponential growth curve from movement onset to offset.

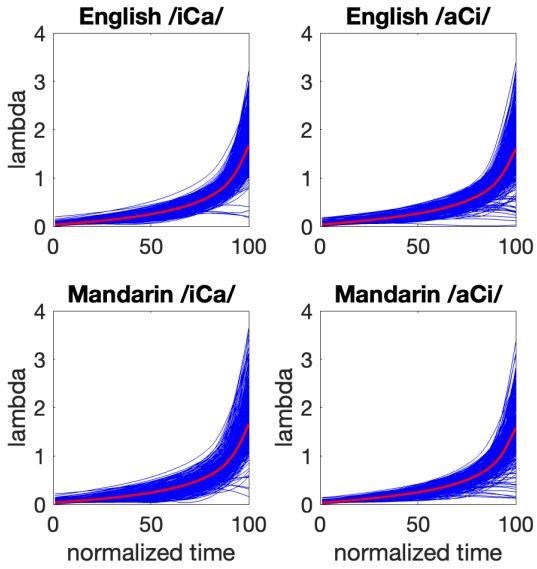


Figure 2: λ *trajectories by language and vowel context. Blue lines show individual trajectories; red lines show average trajectories. Trajectories were normalized to a 100-unit timescale using shape-preserving cubic Hermite interpolation.*

From this observation, it follows that the first time derivative of $\ln(\lambda)$ approximates a constant for each movement, which we call r. To evaluate the robustness of this generalization, a linear regression model was fit to each trajectory of $\ln(\lambda)$ over time. The fits were excellent: overall mean $R^2 = .97$. Moreover, as seen in **Figure 3**, r, the slope of each linear fit, correlates strongly with linguistically relevant measures like duration

(Spearman's $\rho = -.83$, $p < .001$) and kinematic stiffness ($\rho = .82$, $p < .001$).

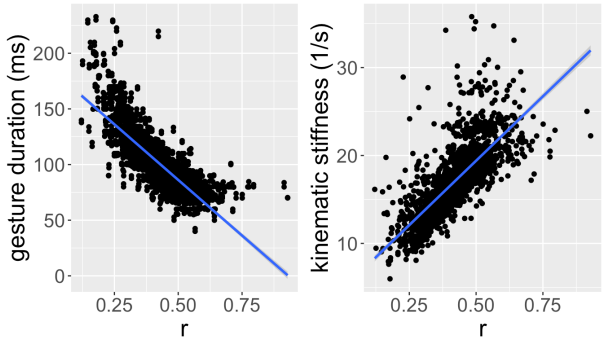


Figure 3: *Correlations between* r *(the slope of a regression line fit to* ln(λ) *and two kinematic variables: gesture duration (left) and kinematic stiffness (right).*

4. Dynamical model

The empirical observation of exponential growth in λ over time can be expressed in the differential equation in (3).

$$\dot{\lambda} = r\lambda \tag{3}$$

Together, the two first order equations in (1) and (3) express a dynamical system of two variables, x and λ. Since λ is defined in (1) as $-\dot{x}/(x - T)$, we can substitute this definition into (3) to derive a single second order equation, eliminating λ. This equation, solved for velocity $\dot{x}$, is shown in (4).

$$\dot{x} = (\ddot{x}/\dot{x} - r)(x - T) \tag{4}$$

(4) has only two parameters, r and T, which can both be inferred from data and have clear interpretations. T corresponds to the spatial target, and r corresponds to movement rapidity, similar to stiffness k in the damped mass-spring model. Moreover, the system is autonomous as it does not reference an extrinsic time variable (Fowler, 1980; Sorensen & Gafos, 2016).

In order to examine the empirical adequacy of (4), we simulated movement trajectories from (4) and compared them to observed trajectories and trajectories simulated from the damped mass-spring model (2). As seen in **Figure 4**, movement trajectories simulated from (4) correspond well with observed trajectories. For instance, peak velocity (corresponding to the zero-crossing in the acceleration curve) occurs 67% of the way through the simulated trajectory, compared to 71% on average (*SD* = 12%) in observed trajectories. For comparison, in the trajectory simulated from the damped mass-spring model, peak velocity occurs 19% of the way through the movement. In both the observed trajectories and the trajectories simulated from our model, the skew in the velocity curve is related to an asymmetry in the acceleration curve: the positive acceleration peak has a smaller magnitude than the negative acceleration peak. In particular, the ratio of the positive peak to the negative peak is 0.51 in the trajectory simulated from our model, compared to 0.83 on average (*SD* = 0.33) in the observed trajectories. In the trajectory simulated from the damped mass-spring model, on the other hand, the positive acceleration peak has a much *greater* magnitude than the negative peak (6.51 times greater).

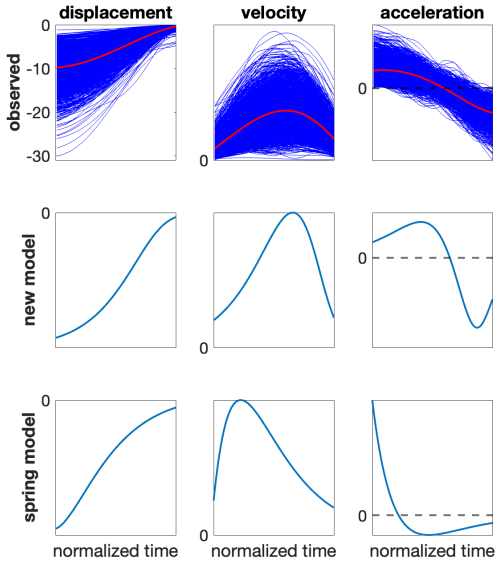


Figure 4: *Displacement (left), velocity (center), and acceleration (right) in real gestures (top), simulated by the proposed model (middle) and simulated by the damped mass-spring model (bottom). All trajectories are demarcated based on a 20% threshold of peak velocity. For both model simulations,* T *= 0 and initial* x *= 10. For the new model simulation,* r *= 10. For the damped mass-spring model simulation,* m *= 1,* b *= 10, and* k *= 25. Trajectories are reversed (multiplied by –1) in order to ease interpretation of velocity and acceleration, and vertical axes are scaled in order to focus on trajectory shapes rather than absolute magnitudes. Dashed horizontal lines indicate acceleration = 0.*

5. Discussion and conclusion

We started from the minimal assumption that articulatory gestures are defined by point attractor dynamics, i.e., a negative relationship between velocity and distance to the target. We formalized this assumption in the differential equation in (1). (1) defines the parameter λ as the negative ratio of velocity to distance to the target, a value which can be measured in articulatory kinematic data. Our investigation of λ trajectories in bilabial constriction movements from 12 English speakers and 12 Mandarin speakers revealed a robust pattern: λ generally follows an exponential growth curve over the course of a movement (**Figure 2**). We incorporated this empirical discovery into the minimal dynamics in (1), deriving (4). Our proposed dynamical system in (4) is both simpler (less parameters) and more empirically adequate than the damped mass-spring model (2). Future work will compare (4) to expanded versions of the damped mass-spring model, i.e., with time-ramped activation (Byrd & Saltzman, 1998; Kröger et al., 1995) or a cubic term (Sorensen & Gafos, 2016). While our model is simpler than those models, a direct comparison of empirical adequacy would be useful in light of the general tradeoff between model simplicity and data fitting.

It is interesting to note that, although (1) is a first order equation—only referencing the first time derivative $\dot{x}$—formalizing the observed temporal variation in λ led to the second order equation in (4). It is not surprising that a second order description is necessary, given that the empirical shapes of velocity curves have proven difficult to capture with first

order dynamics, as described in the Introduction. Although both our model and the damped mass-spring model include an acceleration term, our model captures the shapes of acceleration curves much more closely than the damped mass-spring model, which predicts instantaneous achievement of peak acceleration (**Figure 4**). Our model likely generates more complex acceleration curves because the acceleration term is weighted by velocity, which is itself time-varying. In the damped mass-spring model, on the other hand, the acceleration term is weighted by the constant parameter m.

We have only begun to probe the empirical predictions of our model. For instance, r correlates with peak velocity. In this way, r is similar to k in the damped mass-spring model. However, in our model, the *time to achieve* peak velocity (as a percentage of gesture duration) is stable under variation in r. In the damped mass-spring model, on the other hand, k correlates with both peak velocity and time to achieve peak velocity (e.g., Z. Liu et al., 2022; Mücke et al., 2024). Thus, the damped mass-spring model predicts a negative correlation between peak velocity and time to achieve peak velocity, while our model does not. It would also be valuable to investigate the absolute magnitudes of peak velocity and acceleration, rather than just the shapes of the curves. So far, dynamical modeling work (including this work) has focused on the timing of landmarks, especially peak velocity (e.g., Sorensen & Gafos, 2016). However, the magnitude of, e.g., peak velocity, offers another kinematic dimension to constrain model building, which we have not yet explored in depth.

In future work, we plan to fit the model parameters r and T to data using least squares regression (Iskarous, 2017), rather than estimating them using heuristics. Fitting the model parameters has the potential to shed light on broader theoretical issues, such as intergestural coordination. Preliminary analysis of bilabial release and vowel constriction movements suggests that linear fits to $\ln(\lambda)$ are slightly worse, i.e., mean $R^2 = .91$ and .89, respectively. This is noteworthy because previous work suggests that the timing of target achievement for these two movements (and not consonant constriction) is coordinated (Kramer et al., 2023). It is possible that the fit is worse for these two kinds of movements because their dynamics are coupled in a way that synchronizes target achievement. Thus, model fit may be improved by the addition of a coupling term. This would constitute evidence for target-based gestural coordination (Turk & Shattuck-Hufnagel, 2020), in contrast to onset-based coordination (Nam & Saltzman, 2003).

Model fit for vowel constriction movements and other kinds of (non-labial) consonant movements may also be improved by closer consideration of the nature of targets T. A primary motivation for examining bilabial consonants is that lip aperture is a hypothesized tract variable that corresponds very closely to measurable kinematics. Movements of other articulators like the tongue body are hypothesized to unfold over two tract dimensions: constriction location and constriction degree (e.g., Browman & Goldstein, 1989; Saltzman & Munhall, 1989). In our preliminary analysis of vowel movements, we assumed a single tract variable in 3D space. This allows the target to be straightforwardly estimated from data, but represents a departure from the theoretical proposal of Articulatory Phonology/Task Dynamics. In future work, we plan to develop a method to estimate separate constriction location and constriction degree targets from data. Then, we can examine whether separating movement dynamics into two systems improves the fit of the model. In this way, our model can offer insights into the nature of the tract variables (i.e., x) governing articulatory movement.

6. Acknowledgements

We would like to thank Cherilyn Wang and Ben Kramer for collecting the data and parsing gestural landmarks, and Yuyang Liu for assistance with data processing.

7. References

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, *26*(2), 173–199.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*, 113–133.

Garcia, D. (2010). Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational Statistics and Data Analysis*, *54*(4), 1167–1178.

Iskarous, K. (2017). The relation between the continuous and the discrete: A note on the first principles of speech dynamics. *Journal of Phonetics*, *64*, 8–20.

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. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS)*, 1022–1026.

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.

Liu, Z., Xu, Y., & Hsieh, F. fan. (2022). Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics. *Journal of Phonetics*, *90*.

Mücke, D., Roessig, S., Thies, T., Hermes, A., & Mefferd, A. (2024). Challenges with the kinematic analysis of neurotypical and impaired speech: Measures and models. *Journal of Phonetics*, *102*, 101292.

Nam, H., & Saltzman, E. (2003). A Competitive, Coupled Oscillator Model of Syllable Structure. *Proceedings of the 15th International Congress of Phonetic Sciences*, 2253–2256.

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.

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.

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.

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.

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.

Tiede, M. (2005). *MVIEW: Software for visualization and analysis of concurrently recorded movement data* [Computer software]. Haskins Laboratories.

Turk, A., & Shattuck-Hufnagel, S. (2020). *Speech Timing: Implications for Theories of Phonology, Phonetics, and Speech Motor Control*. Oxford University Press.

Intensity downtrends in Embosi intonation

Yubin Zhang[1], Yijing Lu[1], Annie Rialland[2], Sarah Harper[3], Louis Goldstein[1]
[1]Department of Linguistics, University of Southern California, Los Angeles, USA
[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@ucsf.edu, louisgol@usc.edu

Abstract

Previous studies on utterance-level intonational trends have focused mainly on fundamental frequency (f0), but there is some evidence that other phonetic properties also play a role. The subglottal pressure and intensity variations have been argued to be involved in intonational trends. However, little is known about the dynamics of the intensity trend in intonational contrasts and its relationship with the f0 trend. The current study examines the dynamical patterns of intensity and f0 in declarative and polar question utterances in a Bantu language called Embosi. The results show that both f0 and intensity exhibit initial rising and final lowering, but their kinematic profiles do not always match. To account for the current findings under the framework of articulatory phonology, we propose a pulmonic pressure initiation unit in addition to intonational tone units at the utterance level.

Keywords: f0, intensity, intonation, dynamics

1. Introduction

It is well established that the f0 of an utterance signals sentence intonation. For example, declarative versus question sentence types can be marked by the utterance-level f0 trend. The f0 of declarative intonation in many languages exhibits downtrends, including more global trends like downstep/downdrift and declination, and more localized trend like final lowering (Connell, 2001; Myers, 1996). For polar questions in many languages, the f0 trend manifests as global f0 rising and more localized utterance-final f0 rising (Brunelle et al., 2012; Myers, 1996; Yuan, 2006). Some languages also use a rising-falling f0 contour to signal question intonation, e.g., in Embosi (Rialland & Aborobongui, 2016). In some African languages, the question intonation can be analyzed as the so-called lax prosody, where f0 lowering is a key characteristic (Rialland, 2009). While the f0 aspects of sentence intonation have received wide attention in the literature, linguistic representations of intonation trends have been suggested to be much richer than f0 (Beckman et al., 2010; Vaissière, 2008). For instance, subglottal pressure and intensity have been found to be implicated in intonational trends (Đào & Nguyễn, 2018; Ladefoged, 1968; Yuan, 2006).

Subglottal pressure and intensity are aerodynamical and acoustic variables closely related to respiratory activities, i.e., the initiatory movement of the lungs (Catford, 1997). In the initiation phase of speech production, one or several speech organs called initiators move in a bellow-like or piston-like manner to create various sound sources. Initiatory movement modifies the volume of the vocal tract between the place of articulation and the initiator, leading to positive or negative pressure therein. In pulmonic pressure initiation, the volume of the initiator, i.e., the lungs, is decreased to generate positive pressure in the subglottal vocal tract or the whole vocal tract when there is glottal opening. The dynamics of the pulmonic initiatory movement is determined by elastic recoil force and respiratory muscular effort (Ladefoged, 1968). The elastic recoil force, also called the relaxation pressure, is the sum of the forces from the elastic structures of the abdomen, lungs, and rib cage. Active respiratory muscular effort includes the activities of inhalatory and exhalatory muscles during speech.

Previous studies demonstrate subglottal pressure downtrends in declarative intonation (Fant & Kruckenberg, 2005). For question intonation, Ladefoged (1968) found that the f0 raising in American English polar questions is accompanied by increased subglottal pressure. Previous studies have also found evidence for the intonational trend of intensity (Brunelle et al., 2012; Gelfer et al., 1987; Yuan, 2006). In Swedish declaratives, there are similar rising and falling trends of subglottal pressure and intensity contours (Fant & Kruckenberg, 2005). In Mandarin and Vietnamese, there is evidence that questions exhibit overall larger intensity than declaratives (Brunelle et al., 2012; Đào & Nguyễn, 2018; Yuan, 2006). While subglottal pressure and intensity may also be modulated by variations in supraglottal gestures (Ohala, 1990), there is conflicting evidence for the global trend of supraglottal gestures in declarative utterances (Fougeron & Keating, 1997; Vayra & Fowler, 1992). Evidence for the role of global trends of supraglottal gestures in distinguishing different sentence types is also lacking. Then, variations in subglottal pressure and intensity in intonational trends may originate from pulmonic initiation. As variations in subglottal pressure also affect f0 (Fant & Kruckenberg, 2005; Zhang, 2016), there can be concomitant changes in the intonational trend of f0. Another possibility is that subglottal pressure and intensity patterns reflect the underlying control of the f0 trend per se. Pitch control invovles both laryngeal and respiratory mechanisms (Ohala, 1978; Sundberg, 1992). If pulmonic initiation is used as a lower-level synergistic component of pitch production, subglottal pressure/intensity and f0 may also exhibit similar intonational trends.

While there is evidence for the involvement of subglottal pressure and intensity in sentence intonational contrasts, their dynamical properties remain unknown. Moreover, it is unclear to what extent f0 and pulmonic pressure initiation are independently controlled, and therefore require independent representations in phonology and speech production. Despite the covariation between f0 and intensity, there is some evidence that f0 and intensity aspects may not always parallel each other in intonation. For example, in a respiratory study on utterance-preplanning effects, Fuchs et al. (2015) found that the length of the whole utterance affects breathing parameters like inhalation depth and duration but not initial f0, whereas the length of first prosodic constituent modulates initial f0.

In the current study, we examine the f0 and intensity patterns of sentence intonation in a Bantu language called Embosi (Rialland & Aborobongui, 2016). In Embosi, mora is the tone-bearing unit and there are two lexical tones—high (H) and low (L). For the declarative intonation of Embosi, Rialland & Aborobongui (2016) postulates an utterance-final boundary tone L% to account for the f0 lowering. Their data seem to suggest that f0 begins to be lowered relatively early in

an utterance and it lands hard utterance-finally with a large negative velocity. There also seems to be utterance-initial f0 rising over several moras. The question intonation in Embosi is analyzed as consisting of a global raising component and a rising-falling contour (HL% boundary tone) (Rialland & Aborobongui, 2016). The HL% boundary tone interacts with lexical tones. The H% part of the HL% is attracted to a group of lexical H tones in the last (HH…)L… tone sequence of an utterance. Thus, the L% tone can cause a long-term f0 lowering of lexical L and H tones towards the end of a polar question utterance. For the intensity trend, initial observation suggests similar initial rising and final falling, which remain to be quantitatively verified.

In this study, we examine three alternative hypotheses couched in the theory of articulatory phonology and task dynamics (Browman & Goldstein, 1986; Saltzman & Munhall, 1989). According to the pulmonic pressure initiation hypothesis, a pulmonic initiatory component governed by parameters that vary between declarative and question intonations is the task variable in sentence intonation. Thus, the variation in pulmonic initiatory movement (or subglottal pressure) should lead to parallel acoustic changes in both f0 and intensity dimensions. More specifically, f0 and intensity are expected to have larger initial height, larger initial velocity, and possibly smaller acceleration in polar questions than in declaratives. Alternatively, according to the pitch synergy hypothesis, the task variable for intonational trends is f0. Pulmonic initiation and laryngeal mechanisms are two lower-level components of the synergy for reaching f0 goals. This hypothesis also predicts parallel f0 and intensity trends in declarative versus question intonation. However, it is also possible that the f0 trend does not always parallel the intensity trend. The independent task hypothesis states that intensity and f0 are task variables of independent dynamical systems. This hypothesis predicts some dissociated patterns of f0 and intensity properties.

2. Methods

The audio files for the acoustic analysis were taken from two Embosi corpora (Rialland et al., 2019). Forty-nine declarative-question minimal pairs (49 declarative utterances and 49 polar question utterances) produced by speaker 1 and speaker 2 were taken from the first small corpus for the analysis. The length of the minimal pairs ranges from 6 to 10 moras. A total of 163 declarative and 11 polar question utterances were taken from the second larger corpus. The declarative utterances are 6- to 14-mora utterances produced by speakers 1, 2 and 3, whereas the polar question utterances are 5- to 21-mora utterances produced by speaker 3. The recording was made on a tablet in the field. Participants read the utterances with a fixed distance from the tablet.

In total, we extracted 204 declarative and 60 polar question utterances from the corpora. The large number of utterances ensures randomization of segmental composition of the utterances, minimizing its potential effects on f0 and intensity trends. The analysis focuses on initial events (first three moras) and final events (final three moras). For polar question utterances, the first three moras of an utterance included in the final statistical analysis all occur at or before the H% landing position whereas the final three moras all occur at or after H%. We extracted the mean f0 and intensity of each vocalic moraic interval. The intensity data were mean-normalized for each corpus because the utterances from the second corpus have overall larger intensity than the first one due to recording settings. The f0 and intensity data were analyzed using mixed-effects models. For each acoustic measure, two separate models were fit to analyze the initial and final events. The fixed effects included *Tone* (L versus H), intonation type *IntType* (declarative versus polar question), the linear term for position in the utterance *PosUtt* (initial: 0, 1, 2; final: -2, -1, 0), and the quadratic term for position in the utterance $PosUtt^2$. The linear term *PosUtt* captures velocity whereas $PosUtt^2$ captures acceleration. For interactions, we included *Tone*:*PosUtt* and *Tone*:$PosUtt^2$ to examine tone-specific dynamics. Moreover, we included *IntType*:*PosUtt* and *IntType*:$PosUtt^2$ interactions to test intonational differences in the kinematic profile. For the random effects, we began with the most parsimonious model with random intercepts of *participant* and *item* (utterance) only. Random slopes were also included if they improve model fit.

3. Results

3.1. F0 results

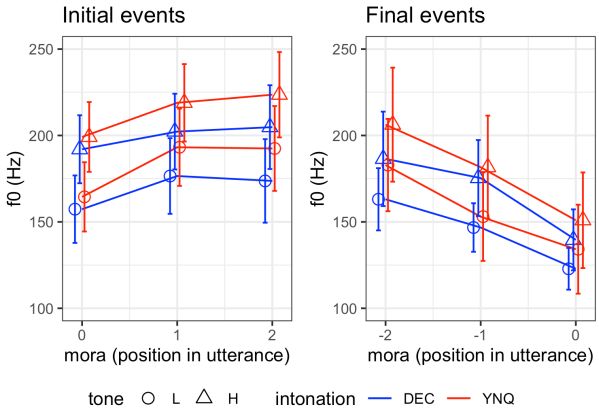


Figure 1. *The f0 patterns of Embosi intonation*

Figure 1 shows the f0 results (mean f0 + standard errors). For the initial event, there is a significant main effect of *Tone* (β = 23.67, $p < 0.001$), confirming higher f0 for the H tone than the L tone. The linear term *PosUtt* (β = 28.74, $p < 0.001$) and the quadratic term $PosUtt^2$ (β = -9.27, $p < 0.001$) also reach significance. The positive estimate for *PosUtt* suggests initial rising with positive velocity. The negative estimate for $PosUtt^2$ suggests negative acceleration, i.e., the positive initial velocity becomes more negative as the utterance unfolds. These two main effects are modulated by *Tone*, as revealed by the significant *Tone*:*PosUtt* (β = -16.13, $p < 0.01$) and *Tone*:$PosUtt^2$ interactions (β = 7.18, $p < 0.01$). The H tone exhibits less positive initial velocity and less negative acceleration than the L tone. Moreover, we found a main effect of *IntType* (β = 7.12, $p < 0.05$), suggesting higher initial f0 height for polar questions than declaratives. The *PosUtt*:*IntType* interaction is also significant, suggesting more positive initial velocity for polar questions than declaratives (β = 13.22, $p < 0.05$).

For final f0 event, the main effect of *Tone* reaches significance (β = 16.74, $p < 0.001$). We also found significant main effects of *PosUtt* (β = -39.70, $p < 0.001$) and $PosUtt^2$ (β = -3.41, $p < 0.01$). The two negative estimates suggest final f0 hard landing, e.g., final f0 lowering with negative velocity and acceleration. The negative velocity becomes increasingly negative towards the end of the utterance. There are significant interactions between *Tone* and *PosUtt* (β = -20.40, $p < 0.01$), and between *Tone* and $PosUtt^2$ (β = 8.54, $p < 0.01$). The lowering of H tones exhibits more negative velocity and acceleration than that of L tones, suggesting more hard landing for the H tones. We also found a significant *PosUtt*:*IntType*

interaction (β = 14.35, p < 0.05). and a significant *PosUtt*^{2}:*IntType* interaction (β = 9.28, p < 0.01). There are less negative velocity and acceleration for the f0 lowering in polar questions than declaratives, suggesting less hard landing for f0 lowering in polar questions.

3.2. Intensity results

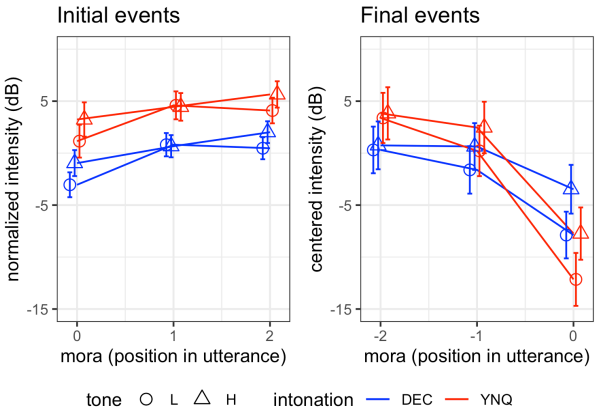


Figure 2. *The intensity patterns of Embosi intonation*

Figure 2 shows the intensity results (predicted mean intensity + standard errors). For the initial event, we found a significant main effect of *Tone* (β = 2.08, p < 0.001). The H tone has larger intensity than the L tone. We also found significant main effects of *PosUtt* (β = 3.59, p < 0.001) and *PosUtt*^{2} (β = -1.05, p < 0.001), suggesting initial intensity rising with positive velocity and negative acceleration. The interactions between *Tone* and *PosUtt* (β = -4.19, p < 0.001), and between the *Tone* and *PosUtt*^{2} (β = 1.96, p < 0.001) also reach significance. The H tone has less positive initial velocity and less negative acceleration than the L tone. Moreover, there is a main effect of *IntType* (β = 4.20, p < 0.001), suggesting larger initial intensity for the polar question than the declarative. We found no interactions between *PosUtt* and *IntType* (β = -0.54, p = 0.63) and between *PosUtt*^{2} and *IntType* (β = 0.12, p = 0.79).

The final event model reveals a significant *Tone* effect (β = 4.41, p < 0.001). We also found significant main effects of *PosUtt* (β = -11.54, p < 0.001) and *PosUtt*^{2} (β = -3.30, p < 0.001), suggesting intensity lowering with negative velocity and acceleration. Intensity lands hard utterance-finally with increasingly negative velocity. We found no evidence for the interactions between *Tone* and *PosUtt* (β = 1.63, p = 0.24), and between *Tone* and *PosUtt*^{2} (β = 0.18, p = 0.79). Moreover, there is a significant main effect of *IntType* (β = -4.27, p = 0.05), suggesting smaller final intensity for the question intonation than the declarative one. The *IntType* effect is mediated by *PosUtt* and *PosUtt*^{2}, as revealed by the significant *PosUtt*:*IntType* interaction (β = -8.51, p < 0.001) and *PosUtt*^{2}:*IntType* interaction (β = -2.42, p < 0.01). There are more negative final velocity and acceleration for intensity lowering in polar questions than declaratives, suggesting more hard landing for intensity lowering in polar questions.

4. Discussion

In the current study, we found that utterance-initially the f0 of an Embosi declarative utterance rises with positive velocity and negative acceleration. Utterance-finally, f0 lands hard, that is, it declines abruptly with negative velocity and acceleration. In polar question intonation, there is similar f0 rising and lowering, but there is larger initial f0 height and initial f0 velocity in polar questions than declaratives. Moreover, f0 lands less hard in polar questions than in declaratives. For tone-specific dynamics, we found that the H tone rises with less positive initial velocity and less negative acceleration than the L tone, whereas the H tone lands harder with more negative velocity and acceleration than the L tone.

The findings are largely consistent with the phonological analysis by Rialland & Aborobongui (2016), but their analysis cannot capture initial f0 rising and more detailed kinematic profiles of the f0 trend in Embosi intonation. In a dynamical framework like articulatory phonology (Browman & Goldstein, 1986; Saltzman & Munhall, 1989), phonological representations and their physical manifestation are unified using dynamical systems. The f0 patterns in Embosi can be captured by global intonational tone gestures coordinated with boundary tone gestures. Differences in sentence intonation can be represented by altering the dynamical gestural units selected and varying the parameter values of the units. More specifically, we hypothesize global intonational tone gestures (H/L) coordinated with boundary tone gestures, like an L% in declaratives and an HL% in polar questions.

The dynamical characterization of these gestures remains to be investigated. The global intonational tone gesture can be hypothesized as an abstract dynamical unit that modulates the f0 targets of individual lexical tones. One possible dynamical system that can account for the initial rising and final hard landing patterns might be a free-fall-style system with parameters like initial height, initial velocity and acceleration. It is likely that H and L tones have different parameter values as we found evidence for tone-specific dynamics in the current study. In polar question intonation, there might be larger values for the parameters like initial f0 height and initial f0 velocity. Moreover, we hypothesize an L% boundary tone at the very end of a declarative utterance and an HL% boundary tone that aligns with the last (HH…)L… lexical tone sequence of a polar question utterance. The boundary tone gestures are hypothesized to have point-attractor dynamics, generating slowing down of the f0 movement towards the target. The less hard f0 landing in polar questions might be explained by the interaction between the soft-landing dynamics of the boundary tone gesture and the hard-landing global intonational dynamics. This is because the L% can have a longer activation interval in questions than in declaratives, extending from the H% to the end of an utterance. Then, if the soft-landing dynamics of the boundary tone unit blends with the hard-landing dynamics of the global intonational tone unit in this long interval, we should observe less hard f0 landing spanning several moras in polar questions. For declarative intonation, the L% might be aligned with the final mora, exerting influences on the final f0 only.

For the intensity trend and its relationship with the f0 trend, the results support neither a pure pitch synergy hypothesis nor a pure pulmonic pressure initiation hypothesis. While intensity and f0 show similar rising and falling patterns, we found evidence for the dissociation between these two properties. There is larger initial f0 velocity for the question intonation than the declarative intonation, but we found no evidence for larger initial intensity velocity in question intonation. Moreover, we found some dissociated landing patterns of f0 and intensity. F0 lands less hard in polar questions than declaratives, but the opposite is true for intensity. For tone-specific results, the H tone has more negative f0 velocity and f0 acceleration than the L tone utterance-finally, but we found no evidence for tone-specific differences in intensity landing. These findings are consistent with the independent task hypothesis, which predicts some non-parallel changes in f0 and intensity (Vaissière, 2008).

Thus, the dimensions of f0 and intensity might be independently controlled in sentence intonation. We propose a pulmonic pressure initiation unit in addition to intonational tone units to account for the findings. We hypothesize that this dynamical unit is a respiratory variable, like lung volume, that controls pulmonic pressure initiation (Catford, 1997). The dynamics of lung volume deflation is contingent on two lower-level articulatory components. The first one is a global trajectory caused by elastic recoil force and the second one is the self-imposed respiratory muscular effort during speech exhalation (Catford, 1997; Ladefoged, 1968). The simulation study by Zhang (2016) reveals that without additional self-imposed muscular effort during speech exhalation, the utterance-level dynamics of lung deflation and the resulting subglottal pressure exhibit a soft-landing pattern like an exponential decay. The exact dynamical properties, like the rate of lung deflation, are affected by initial lung volume. Moreover, when the dynamics of additional muscular effort is considered, different dynamics of lung deflation and subglottal pressure can further emerge.

In Embosi, the pulmonic pressure initiation unit is hypothesized to exhibit some hard landing dynamics as with the global intonational tone unit. Since f0 and pulmonic pressure initiation are hypothesized to be independently controlled, the dissociation between f0 and intensity patterns can be readily accounted for. For example, the independently controlled pulmonic pressure initiation unit may only have an increased value for the initial lung volume parameter in polar question intonation but not for initial lung initiator velocity. Additionally, the landing pattern of intensity is hypothesized to be caused primarily by the pulmonic pressure initiation unit, that is, there might be more negative acceleration for pulmonic pressure initiation in polar questions. The f0 trajectory can show distinct dynamical properties due to the interaction among intonational tone gestures.

It should be noted that while the current findings are consistent with the independent task hypothesis, it does not preclude the possibility that tone production can also engage pulmonic initiation as a synergistic component of pitch control. Indeed, we found that the H tone has larger intensity than the L tone, suggesting a possible role of using pulmonic initiatory mechanisms for achieving H tones. Different types of mechanisms underlying f0 and intensity control may co-exist in the production of tonal and intonational contrasts. A computational model that can account for the observed patterns remains to be developed in future research. Respiratory studies are also needed to reveal a fuller picture of respiratory and tonal f0 dynamics in intonational contrasts.

5. References

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*. https://doi.org/10.1093/acprof:oso/9780199249633.003.0002

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–252. https://doi.org/10.1017/s0952675700000658

Brunelle, M., Ha, K. P., & Grice, M. (2012). Intonation in Northern Vietnamese. In *Linguistic Review* (Vol. 29, pp. 3–36). https://doi.org/10.1515/tlr-2012-0002

Catford, J. C. (1997). *Fundamental problems in phonetics.* Edinburgh University Press.

Connell, B. (2001). Downdrift, downstep, and declination. In *Proceedings of the TAPS (Typology of African Prosodic Systems Workshop) May 18-20, 2001* (pp. 1–8).

Đào, Đ. M., & Nguyễn, A. T. T. (2018). Acoustic correlates of statement and question intonation in Southern Vietnamese. *Journal of the Southeast Asian Linguistics Society*, *11*(2), 19–41.

Fant, G., & Kruckenberg, A. (2005). Covariation of subglottal pressure, F0 and intensity. In *9th European Conference on Speech Communication and Technology* (pp. 1061–1064). https://doi.org/10.21437/interspeech.2005-425

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728–3740. https://doi.org/10.1121/1.418332

Fuchs, S., Petrone, C., Rochet-Capellan, A., Reichel, U. D., & Koenig, L. L. (2015). Assessing respiratory contributions to f0 declination in German across varying speech tasks and respiratory demands. *Journal of Phonetics*, *52*, 35–45. https://doi.org/10.1016/j.wocn.2015.04.002

Gelfer, C. E., Harris, K. S., & Baer, T. (1987). Controlled variables in sentence intonation. In *Laryngeal function in phonation and respiration* (pp. 422–435).

Ladefoged, P. (1968). Linguistic aspects of respiratory phenomena. *Annals of the New York Academy of Sciences*, *155*(1), 141–151. https://doi.org/10.1111/j.1749-6632.1968.tb56758.x

Myers, S. (1996). Boundary tones and the phonetic implementation of tone in Chichewa. *Studies in African Linguistics*, *25*(1), 29–60. https://doi.org/10.32473/sal.v25i1.107403

Ohala, J. J. (1978). Production of tone. In *Tone: a linguistic survey* (pp. 5–39). Academic Press. https://doi.org/10.1016/b978-0-12-267350-4.50006-6

Ohala, J. J. (1990). Respiratory activity in speech. In *Speech Production and Speech Modelling* (pp. 23–53). https://doi.org/10.1007/978-94-009-2037-8_2

Rialland, A. (2009). The African lax question prosody: Its realisation and geographical distribution. *Lingua*, *119*(6), 928–949. https://doi.org/10.1016/j.lingua.2007.09.014

Rialland, A., & Aborobongui, M. E. (2016). How intonations interact with tones in Embosi (Bantu C25), a two-tone language without downdrift. In *Intonation in African Tone Languages* (pp. 195–222). https://doi.org/10.1515/9783110503524-007

Rialland, A., Adda-Decker, M., Kouarata, G. N., Adda, G., Besacier, L., Lamel, L., … Cooper-Leavitt, J. (2019). Parallel corpora in Mboshi (Bantu C25, Congo-Brazzaville). In *LREC 2018 - 11th International Conference on Language Resources and Evaluation* (pp. 4272–4276).

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

Sundberg, J. (1992). Breathing behavior during singing. *Dept. for Speech, Music and Hearing Quarterly Progress and Status Report*, *33*(1), 49–64.

Vaissière, J. (2008). Perception of intonation. In *The Handbook of Speech Perception* (pp. 236–263). https://doi.org/10.1002/9780470757024.ch10

Vayra, M., & Fowler, C. A. (1992). Declination of supralaryngeal gestures in spoken Italian. *Phonetica*, *49*(1), 48–59. https://doi.org/10.1159/000261902

Yuan, J. (2006). Mechanisms of question intonation in Mandarin. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4274 LNAI, pp. 19–30). https://doi.org/10.1007/11939993_7

Zhang, Z. (2016). Respiratory laryngeal coordination in airflow conservation and reduction of respiratory effort of phonation. *Journal of Voice*, *30*(6), 760.e7-760.e13. https://doi.org/10.1016/j.jvoice.2015.09.015

Lingual and epilaryngeal articulation of vowels in Mundabli

Matthew Faytak[1], Mariana Quintana Godoy[1], Tianle Yang[1]

[1]*University at Buffalo, USA*

{faytak, marianaq, tianleya} @buffalo.edu

Abstract

The vowel system of Mundabli (Yemne-Kimbi, Cameroon) is rich in contrasts involving lower vocal tract activity. In this study, we aim to characterize the acoustics and articulation of the three sets of vowels in the language, which we refer to as pharyngeal, plain, and lax. Acoustic time series data reveal that pharyngealization raises F1 and lowers F2 and F3; it also conditions tense or creaky voice quality relative to the plain and lax vowels. Ultrasound data suggest that these acoustic properties can typically be attributed to a lower pharyngeal or epilaryngeal constriction. The data also suggest that the lax and plain vowels may exhibit an advanced tongue root contrast. Variation in the articulatory implementation of pharyngealization observed in the ultrasound data is discussed.

Keywords: speech production, ultrasound tongue imaging, laryngeal articulator, pharyngealization, advanced tongue root

1. Introduction

Mundabli (ISO 639-3 [boe]) is a Yemne-Kimbi language spoken in the Lower Fungom region of northwestern Cameroon by no more than 800 inhabitants of a single village of the same name (Good et al. 2011; Voll 2017). Mundabli's vowel system features ten monophthongs /i ɪ e ɛ ɨ a u ʊ o ɔ/ plus six monophthongs which have been described as pharyngealized /iˤ eˤ ɨˤ aˤ uˤ oˤ/ (Voll 2017). The paired arrangement of the front and back non-pharyngealized vowels suggests two sets contrasting in height or tongue root advancement, here referred to as PLAIN /i e u o/ and LAX /ɪ ɛ ʊ ɔ/.

The PHARYNGEAL vowels are unusual both in Mundabli's local area and more broadly cross-linguistically. They have developed recently in the language's history from vowels formerly followed by coda **k* or **ʔ*. Fieldwork by the first author has revealed that a closely related neighboring language, Mufu, has /k/ or /ʔ/ in cognate lexical items, e.g. Mufu [bàʔ], Mundabli [bàˤ] 'scar'; Mufu [cōk], Mundabli [tsōˤ] 'banana'; Mufu [dàk], Mundabli [dèˤ] 'place'.

Morphophonological alternations between plain and pharyngeal vowels suggest a phonological organization of the vowels into a system summarized in **Table 1**. For instance, pharyngealization marks imperfective aspect on many open-syllable verb stems, e.g. [fì] 'press.PERF' *vs.* [fìˤ] 'press.IPFV'; [bú] 'give birth.PERF' *vs.* [búˤ] 'give birth.IPFV'. This alternation is a trace of an imperfective suffix **-k(ə)* which is still overtly realized in Mufu as *-k* or *-ʔ*; compare Mufu [fjàk] 'press.IPFV', [búk] 'give birth.IPFV'. We refer the reader to Voll (2017) for further reading on the relationships among the Mundabli vowel sets.

Mundabli is notably rich in contrasts based on lower vocal tract activity, in that both the lax and pharyngeal vowels may make use of the epilaryngeal tube as an articulator. Pharyngeal consonants and pharyngealized vowels are known to involve strong constriction of the epilaryngeal tube (Catford 1983; Arkhipov et al. 2019). On the other hand, the lax vowel set may exhibit retracted tongue root (RTR), and the plain vowel set advanced tongue root (ATR), a common arrangement elsewhere in West Africa (Casali 2008). RTR vowels have *also* been identified as making use of epilaryngeal constriction to distinguish themselves from ATR vowels (Esling 2005; Edmondson et al. 2007; Starwalt 2008).

It is unclear how Mundabli speakers would organize lingual and epilaryngeal articulation in its vowels to accommodate a three-way lower vocal tract activity distinction, since such a situation is (at the least) very rare in the world's languages. As such, this exploratory study aims to clarify the acoustic differences among Mundabli's pharyngeal, plain, and lax vowel sets, and the lingual and epilaryngeal articulatory basis of these distinctions. In particular, we hope to clarify the articulatory nature of the pharyngeal and lax vowels, and how they differ from the plain vowels and each other.

Table 1: *A possible phonological organization of the Mundabli monophthongal vowels.*

Vowel set	Plain	Pharyngeal	Lax
I	i	iˤ	ɪ
E	e	eˤ	ɛ
II	ɨ	ɨˤ	
A	a	aˤ	
U	u	uˤ	ʊ
O	o	oˤ	ɔ

2. Methods

Time-aligned acoustic and ultrasound data were collected from 15 Mundabli speakers (7M 8F, mean age 31.9, SD 8.01) in 2022 and 2023 in Douala, Cameroon using Articulate Assistant Advanced (v221.2.0). Audio was recorded at a sampling rate of 22.05 kHz using a Røde NTG2 supercardioid condenser microphone mounted on a tabletop tripod; the signal was digitized using a Scarlett Solo 2i2 USB audio interface. Ultrasound video was recorded using a Telemed MicrUs and an MC4 micro-convex probe secured by an Articulate Instruments UltraFit headset (Spreafico, Pucher, and Matosova 2018); recordings were made at a rate of 82.1 Hz in a 101.2° field of view.

Stimuli were a set of 32 open-syllable monosyllabic words containing all 16 Mundabli vowels (two word types per vowel) read in one of three frame sentences, verbally prompted by the first author. The frame sentences were varied to provide a syntactically appropriate frame for each target word. Nouns were presented in the frame ⁿkáˤn _ 'I have (a) (noun)"; verbs were

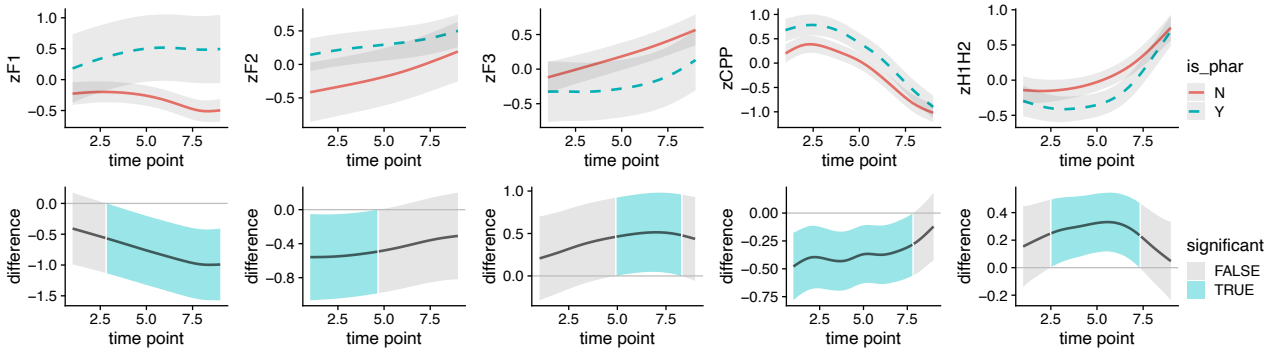


Figure 1: *GAMM smooths and difference smooths for z-scored F1, F2, F3, CPP, and H1*-H2* over vowels' normalized duration.*

presented one of two frames, ⁿfãˆ áˆ ⁿ _ 'I am (verb)-ing' for imperfective forms of verbs, which invariably take a low-toned and prenasalized form in this context, and ⁿfá _ 'I have (verb)-ed' The varied frame sentences were needed, in part, to elicit the correct aspect-inflected form of verbs which are pharyngealized only in the imperfective. During recording, speakers repeated the frame with the embedded target between four and six times (typically five).

The final acoustic data set contains 4,726 vowel tokens across all 15 speakers. Acoustic measures were extracted from this data using PraatSauce (Kirby 2019) at nine evenly spaced time points across vowels' durations. Formant frequencies (F1-3) were extracted, and the amplitude difference between the first and second harmonics (H1-H2) and cepstral peak prominence (CPP) were extracted as measures of voice quality. All measures were *z*-scored. After removing outliers more than ±2SD from the mean for each measure, 37,092 samples remained (from roughly 4,121 vowel tokens). AR1 GAMMs were carried out for each acoustic measure using the *mgcv* package in R (Wood 2023), with factor smooths for vowel class (pharyngealized *vs* pooled lax and plain) and random smooths for speaker and word.

The co-collected ultrasound data was analyzed for six of the 15 speakers (3M 3F, mean age 30.0, SD 7.92). Tongue surface contours were segmented from ultrasound video using the DeepLabCut model implemented in Articulate Assistant Advanced (v221.2.0). DeepLabCut contours were converted to fan splines and trimmed of any knots not originally estimated by the DeepLabCut model; contours at vowel midpoint were extracted and converted to polar coordinates. Smoothing-spline ANOVAs (SSANOVAs) by vowel group and category (pharyngeal *vs.* plain *vs.* lax) were carried out using the *gss* package in R (Gu 2023).

3. Results

3.1. Acoustic measures

GAMM smooths and difference plots for formant data are shown in the first three columns of **Figure 1**; difference smooths are located below each measure's smoothed estimates. As a group, pharyngealized vowels exhibit raised F1, raised F2, and lowered F3 (Figure 1B) relative to the plain and lax vowels. The difference in F1 is largest, and increases in size towards the end of the vowel. This coincides with the development of a smaller, significant difference in F3 during the last half of the vowel. By contrast, the F2 difference reaches significance only during the first half of the vowel and diminishes later in the vowel.

GAMM smooths and difference plots for voice quality data are shown in the two rightmost columns in **Figure 1**. Pharyngealized vowels exhibit elevated CPP and reduced H1*-H2* relative to their plain and lax counterparts, suggesting that pharyngealized vowels are characterized by relatively tense or creaky phonation. Values for both voice quality measures converge towards the end of vowel duration in a direction that suggests that all vowels end in breathy phonation; this is likely a reflection of the prepausal devoicing of the end of all target words. Phrase-final devoicing is a tendency noted independently by Voll (2017, p. 32).

3.2. Ultrasound

Data for six of fifteens speakers are analyzed here, with SSANOVAs carried out separately for each speaker. For reasons of space, we present results pooled across all vowels by speaker, then provide by-vowel group breakdowns for two representative speakers.

We first focus on articulatory differences between the pharyngeal and non-pharyngeal vowels (lax and plain vowels as a group). For the pharyngeal vowels, all speakers exhibit substantial lowering of the posterior tongue dorsum and bunching and fronting of the anterior dorsum and blade (**Figure 2**). Some speakers tend toward a double-bunched shape, particularly speaker F1 (**Figure 3**). The pharyngeal-non pharyngeal difference is reduced somewhat in the front vowels, particularly in the I group (**Figures 3, 4**). For four of the six speakers (F1, M1, M2, and M3) the lower portion of the tongue root shows additional retraction relative to the non-pharyngeal vowels. This difference obtains across all vowel groups, though less so for the back vowel groups (U and O), as shown for speaker F1 in **Figure 3**. Speaker F3 exhibits no differences in the position of the lower tongue root, but her pattern of articulation otherwise resembles the aforementioned four speakers.

Speaker F2 diverges from the majority pattern, with the entire tongue root fronted during production of pharyngealized vowels. This pattern also holds in most of the individual vowel groups as shown in **Figure 4**, though the I group exhibits no clear distinction in root frontness and the E group exhibits the majority pattern by which the root is retracted. Speaker F2's articulation of the pharyngeal vowel set could thus be characterized as involving fronting of the entire tongue and a bunched configuration, rather than the simultaneous fronting/bunching and lower root constriction which characterizes four of the five remaining speakers.

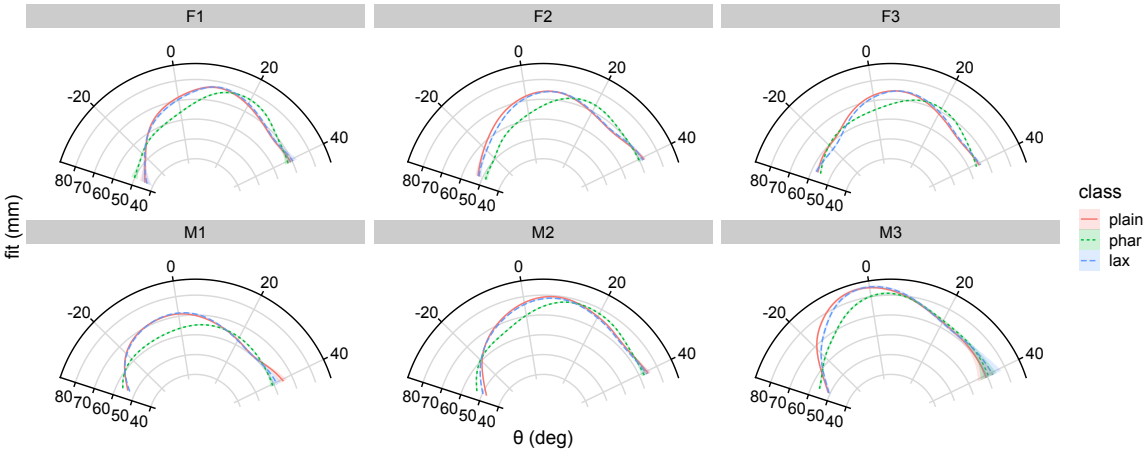


Figure 2: *SSANOVA splines at vowel midpoint for all speakers, pooling across vowel groups.*

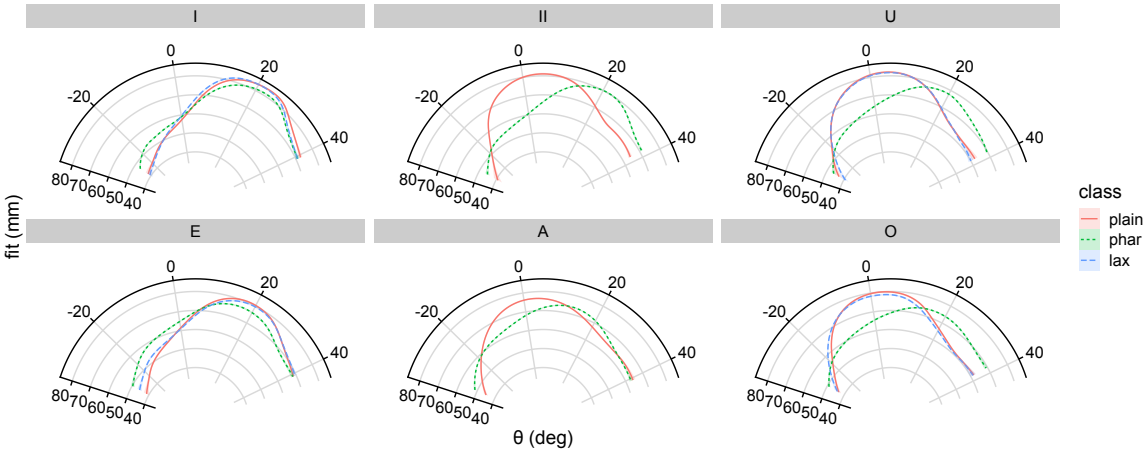


Figure 3: *SSANOVA splines at vowel midpoint by vowel group for speaker F1, as an example of the majority pattern of root retraction in pharyngealized vowels.*

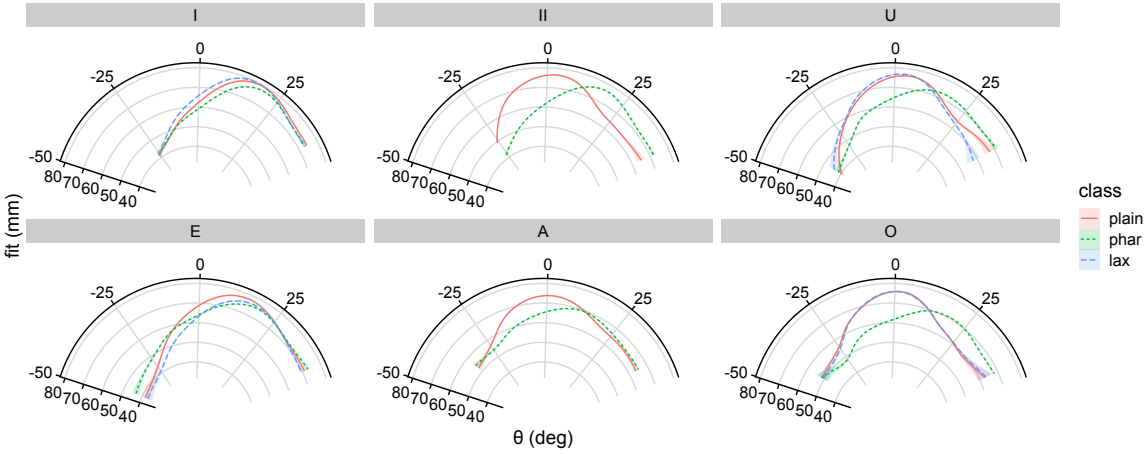


Figure 4: *SSANOVA splines at vowel midpoint by vowel group for speaker F2, who generally exhibits a distinct pattern of tongue fronting in pharyngealized vowels.*

We now turn to differences in articulation between the lax and plain vowel sets. Unpooling the vowels as in **Figures 3, 4**, for both speakers F1 and F2, a majority of the data suggests slight tongue root retraction for lax /ɪ ɛ ʊ ɔ/, relative to plain /i e u o/, with little consistent involvement of dorsum height differences. Both speakers exhibit reversals of this relationship (F1's /u-ʊ/ pair, F2's /e-ɛ/ pair), but these may be due to gaps in the lexicon which required the use of stimuli with lingual onset consonants in the E and U vowel groups (i.e. gbɛ́ 'wind', gbʊ̄ 'fall').

Finally, we compare the lax and pharyngeal vowels for speaker F2 in **Figure 3**, whose pharyngeal vowels involve root retraction. The root retraction observed for the lax set is less extreme than for the pharyngeal set; the lax set's root retraction also recruits tongue dorsum lowering to a lesser extent.

4. Discussion and conclusion

The articulatory and acoustic data suggest that the pharyngeal vowels involve an epilaryngeal constriction for most speakers, a maneuver readily distinguished from lower vocal tract articulations often called "pharyngealization", such as emphasis or uvularization (Evans et al. 2016; al-Tamimi 2017). Lower pharyngeal or epilaryngeal constriction is clearly suggested by both the F2-raising effect seen in the pharyngeal vowels (as opposed to F2-lowering for uvularization) and the characteristic double-bunching also observed in languages with lower pharyngeal constrictions (Catford 1983; Arkhipov et al. 2019). The involvement of tense or creaky phonation in pharyngeal vowels also specifically implicates the laryngeal articulator, as constriction of the epilarynx is known to yield non-modal phonation (Edmondson et al. 2007; Moisik, Czaykowska-Higgins, and Esling 2021).

The "lax" and plain vowels appear to exhibit ±ATR differences. This finding suggests that Mundabli may exhibit two articulatory mechanisms for the lax and pharyngealized vowels which involve different types or degrees of epilaryngeal constriction. The Mundabli pharyngealized vowels often show dorsum lowering and "bunching" around the level of the hyoid bone, a configuration which Esling (2005) describes as characteristic of the most epilaryngeally constricted sounds; they especially resemble pharyngeal approximant consonants such as [ʕ]. The lax vowels are lightly retracted by comparison, without a great deal of tongue dorsum lowering in support of this goal. The lax vowels thus more closely resemble vocalic ATR or registral differences as described in a number of previous works (Edmondson et al. 2007; Esling 2005). Further analysis of more speakers' articulations is needed within each vowel group to confirm the robustness of this finding. More detailed analysis of the acoustic data to examine the voice quality of the lax vowels is also merited, given that ATR distinctions are often associated with nonmodal phonation to some degree (Casali 2008; Starwalt 2008; Akinbo et al. 2023).

Speaker F2 appears to use fronted or singly bunched, rather than double-bunched, tongue shapes for the pharyngeal vowels, with no obvious lower pharyngeal constriction. The lack of epilaryngeal activity for this speaker is also suggested by the absence of CPP and H1*-H2* differences between the pharyngealized and non-pharyngealized vowels for this speaker; on the other hand, her formant data are comparable to that of other speakers. For speaker F2, the contrast between the pharyngeal vowels and the other vowel sets may have shifted to rely more on the associated formant frequency differences, possibly due to a misattribution of the acoustic effects of double-bunching to a fronted, singly bunched tongue position. Three other speakers in the data set whose ultrasound data were not analyzed here may exhibit similar variants to speaker F2's; further analysis may shed light on variation across the larger population of Mundabli speakers.

5. Acknowledgements

Yùŋ kɛ́n mòˀmòˀ to the Mundabli community members for your contributions as language experts. We also thank Alan Wrench, Bowei Shao, and Jalal al-Tamimi for helpful comments, Connor Quimby for annotation of Mufu data, and Ikom Christopher, Jeff Good, and Pierpaolo DiCarlo for logistical support.

6. References

Akinbo, Samuel, Avery Ozburn, Gerald Nweya, and Douglas Pulleyblank (2023). "Eleven vowels of Imilike Igbo including ATR and RTR schwa". In: *JIPA*, pp. 1–24.

Arkhipov, A, M Daniel, O Belyaev, G Moroz, and JH Esling (2019). "A reinterpretation of lower-vocal-tract articulations in Caucasian languages". In: *Proc ICPhS 19, Melbourne*, pp. 5–9.

Casali, Roderic F (2008). "ATR harmony in African languages". In: *Lang and Ling Compass* 2.3, pp. 496–549.

Catford, John C (1983). "Pharyngeal and laryngeal sounds in Caucasian languages". In: *Vocal fold physiology: Contemporary research and clinical issues*, pp. 344–350.

Edmondson, Jerold A, Cécile M Padayodi, Zeki Majeed Hassan, and John H Esling (2007). "The laryngeal articulator: Source and resonator". In: *Proc ICPhS 16, Saarbrücken*. Vol. 3, pp. 2065–2068.

Esling, John H (2005). "There are no back vowels: The larygeal articulator model". In: *Canadian Journal of Linguistics/Revue canadienne de linguistique* 50.1-4, pp. 13–44.

Evans, Jonathan P, Jackson T-S Sun, Chenhao Chiu, and Michelle Liou (2016). "Uvular approximation as an articulatory vowel feature". In: *JIPA* 46.1, pp. 1–31.

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.

Gu, Chong (2023). *Package 'gss'*. R package version 2.2.7.

Kirby, James (2019). *praatsauce*. GitHub repository. URL: `https://github.com/kirbyj/praatsauce`.

Moisik, Scott R, Ewa Czaykowska-Higgins, and John H Esling (2021). "Phonological potentials and the lower vocal tract". In: *JIPA* 51.1, pp. 1–35.

Spreafico, Lorenzo, Michael Pucher, and Anna Matosova (2018). "UltraFit: A speaker-friendly headset for ultrasound recordings in speech science". In: *Proc Interspeech 2018, Hyderabad*. ISCA, pp. 1517–1520.

Starwalt, Coleen G. A. (2008). "The acoustic correlates of ATR harmony in seven-and nine-vowel African languages: A phonetic inquiry into phonological structure". PhD thesis. U Texas Arlington.

al-Tamimi, Jalal (2017). "Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: Implications for formal representations". In: *LabPhon* 8.

Voll, Rebecca (2017). "A grammar of Mundabli: a Bantoid (Yemne-Kimbi) language of Cameroon". PhD thesis. U Leiden.

Wood, Simon (2023). *Package 'mgcv'*. R package version v1.8.42.

Spatiotemporal Coupling of the Jaw and Lower Lip: Comparing Talkers with Parkinson's Disease and Amyotrophic Lateral Sclerosis

Mili Kuruvilla-Dugdale[1], Antje Mefferd[2]

[1]*University of Iowa*
[2]*Vanderbilt University Medical Center*
mkuruvilladugdale@uiowa.edu, antje.mefferd@vumc.org

Abstract

This study sought to determine the differences in temporal coupling between the lower lip and jaw for two gestures i.e., the vowel /ʌ/ and the labiodental fricative /f/ in the word 'muffin.' Because articulatory timing can be disrupted by impairments to the basal ganglia and its role in intrinsic timing and the dynamic state of the articulatory system, interarticulator timing was compared between tallkers with amyotrphic lateral sclerosis (ALS) and Parkinson's disease (PD) relative to healthy controls. Electromagnetic articulography was used to record lower lip and jaw movements from six talkers with ALS, nine with PD, and 10 healthy controls. Lag values were obtained by subtracting the timepoint of the lower lip from the timepoint of the jaw for each gesture based on the timepoints of the positional minima for the jaw and lower lip during /ʌ/ and the positional maxima for the jaw and lower lip during /f/. Absolute lag times, percent lag (relative to total word duration), and coefficient of variation (CoV) values were compared between the three groups, as were effect sizes. Our results show a trend towards greater interarticulator timing differences (i.e., less temporal coupling) in talkers with ALS, whereas talkers with PD showed similar timing patterns to healthy controls. CoV values tended to be lower in the clinicial groups, with the ALS group showing more consistent lag times than even the PD group. Although preliminary, these results provide evidence of a mismatch between intrinsic timing and the physical state of the articulators in talkers with ALS. Despite basal ganglia pathology, the relative timing patterns among articulators appeared to be intact in talkers with PD.

Keywords: interarticulator timing, dysarthria, articulator coupling

1. Introduction

Intelligible speech requires careful timing of articulatory movements. Within the articulatory phonology framework, a high level of coupling is expected among articulators that form a gesture (Saltzman & Munhall, 1989). Within gestures, the interarticulator timing is supposed to follow a specific order. For example, for lip closing, the peak velocities of the lips are known to lead those of the jaw (e.g., Gracco, 1988). Similarly, when reaching the target position for a consonant, the jaw usually follows the tongue tip (Mooshammer et al., 2006). However, it is currently unknown if the inter-articulatory timing patterns seen in healthy, mature talkers are maintained by talkers with neurological conditions. It is also unknown to what extent disruptions to interarticulatory timing patterns may differ across talkers with different pathophysiologies within the speech motor system.

The temporal coupling among articulators is driven by a central clock linked to several cortical and subcortical areas such as the basal ganglia and cerebellum (Grahn, 2009; Konoike et al., 2012). This intrinsic clock is thought to be comprised of multiple oscillators at the gestural and suprasegmental levels that shape the motor plan to insert temporal with linguistic information (Saltzman et al., 2008; Windmann et al., 2015). Among the neural structures, there are differences in the roles of the basal ganglia and the cerebellum. The basal ganglia are engaged in the processing of attention-based, longer temporal intervals whereas the cerebellum is concerned with automatic, shorter, and event-based temporal processing (Harrington et al., 1998; Meck, 2005). The output of the central oscillators serves as input to the articulators and interacts with their physical state (e.g., stiffness) to shape the surface movement patterns. Impairments to both the central clock and dynamic state of the articulatory system can disrupt articulatory timing (Rong & Heidrick, 2022).

In individuals with amyotrophic lateral sclerosis (ALS), the articulators undergo significant morphological changes with disease progression, which alters their intrinsic properties and functional capacity as evidenced by reduced force generation, and slow and reduced movements (Lee & Bell, 2018; Shellikeri et al., 2016). Particularly the tongue is disproportionally more affected by the disease than the lips and jaw (e.g., Langmore & Lehman, 1994; DePaul et al., 1988). Therefore, the timing information generated by the central clock presumably interferes with the dynamic state of the articulators. In other words, there is likely a mismatch between the designated time determined by the linguistic event, and the physical properties of the articulators. Given the differential impairment of the articulators, the jaw is thought to become a primary articulator moving the tongue and perhaps also the lower lip more passively. This may result in more synchronized movements of the jaw and tongue or lower lip.

Evidence for the role of the basal ganglia in representing temporal information comes from multiple sources, including studies on Parkinson's disease (PD) and Huntington's disease. These studies report interval timing dysfunction (Malapani et al., 1998). Functional magnetic resonance imaging studies have also found that the striatum is activated by tasks that involve interval information processing durations (Tanaka et al., 2007). There is some debate about the exact role of the basal ganglia as some studies have shown that administration of dopamine agonists increases the speed of the internal clock (Maricq & Church, 1983; MacDonald & Meck, 2005); while others could not demonstrate such an effect (Balcı et al., 2008). Yet, others reported an opposite effect suggesting that increased dopamine levels might decrease the speed of timekeeping (Lake & Meck, 2013).

Despite mixed findings, there is consensus that basal ganglia disorders like PD disrupt temporal articulatory patterns. However, timing patterns between the jaw and the primary articulators (e.g., tongue, lower lip) have not been studied in these talkers. One study examined the intergestural timing patterns in talkers with essential tremor (Hermes et al., 2019), a neurological condition also associated with basal ganglia pathology. The coordination pattern between the tongue tip and

tongue back for simple CV syllables were similar between the clinical group and healthy control speakers. However, during CCV syllables, which are thought to be phonetically more challenging, the coordination patterns of the lip, tongue tip, and tongue back significantly differed between the two groups, and these deviant patterns further degraded during deep brain stimulation. That is, participants in the clinical group activated gestures for both consonants and the vowel all at once. They also lengthened the prevocalic consonant considerably indicating their inability to adequately sequence these movements during the CCV gesture. However, it is unclear to what extent these findings translate to talkers with PD and how they relate to inter-articulatory timing patterns of the jaw and a primary articulator (e.g., tongue, lower lip) within a gesture.

To address the current gap in the literature on interarticulator timing patterns in talkers with dysarthria, the current study examined the timing between the lower lip and jaw during two gestures (open vowel /ʌ/ and labiodental fricative /f/) in talkers with ALS and PD. Specifically, as a first step, this study aimed to determine the extent to which jaw and lower lip are coupled (synchronized) in these talkers. Furthermore, we sought to determine how stable (consistent) these timing patterns were across trials. Because talkers with ALS activate the jaw more during speech than their healthy peers, and likely rely more heavily on the jaw as a primary articulator to achieve the desired vocal tract configuration, we expect more synchronized timing pattern (smaller lag times relative to controls) for the lower lip and the jaw in these talkers. Based on the consensus that their basal ganglia pathology disrupts temporal patterns in talkers PD, deviant interarticulatory pattern may be observable in these talkers; however, prediction about the specific direction (more or less synchronized than controls) could not be made. However, it should be noted that the basal ganglia pathology of talkers with PD is conceptualized to affect absolute timing patterns such as speech rate or segment durations rather than relative timing patterns. In that case, interarticulatory timing patterns of the lower lip and jaw of talker with PD may be similar to those of controls. Finally, we did not formulate specific hypotheses for the trial-to-trial variability of lag times but potential group differences will be explored.

2. Methods

This study was approved by the Institutional Review Board at the University of Missouri and Vanderbilt University Medical Center. All participants provided consent prior to data collection and were compensated for their time.

2.1. Participants

Participants belonging to three groups, namely ALS, PD, and healthy controls were included in the study. So far, we have collected data from six talkers with ALS (6 males, M_{age}=65.33, *SD*=9.63), nine with PD (5 females, 4 males, M_{age}=70.44, *SD*=5.62), and 10 controls (9 females, 1 male, M_{age}= 56.77, *SD*=5.72). Talkers with ALS and PD ranged in their dysarthria severity from mild to moderate-severe. All participants were monolingual native speakers of American English.

2.2. Kinematic Data Collection

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 and lower lip (lowering for the open vowel /a/ and raising for the labiodental fricative /f/). Articulatory kinematic data from all but one participant were collected using the Wave Speech Research System (NDI, Waterloo, Ontario, Canada) and data from one participant with ALS was collected with the AG501 (Carstens Medizinelektronik, GmbH, Nelkenweg, Germany). To record speech kinematics, small sensors were affixed along the mid-sagittal plane of the articulators (i.e., tongue tip, jaw, lips). The tongue tip sensor was placed at 1 cm from the tip; lower lip sensor was placed on the vermillion border, and the jaw center sensor was affixed to the lower gum below the central incisors. A head reference sensor recorded the head movements. Kinematic data were corrected for head movements and rotated into a head-based coordinate system using software provided by NDI. For recordings with the AG501, participants were asked to hold a bite plate with three additional sensors in their mouth. This recording was later used to transpose the kinematic data into a head-based coordinate system with the origin located just anterior to the jaw center sensor (Mefferd, 2017). This biteplate correction creates a head-based coordinate system that is comparable to that of the Wave system.

The sampling rate for the AG501 was 1250 Hz, further down sampled to 250 Hz, and for the Wave system, the sampling rate was 400 Hz. The audio signal was synchronized with the kinematic data and was sampled at 48,000 Hz and 22,000 Hz for the AG501 and the Wave systems, respectively. All kinematic data were low pass filtered at 15Hz. For this study, only the kinematic data of the lower lip and the jaw were used.

2.3. Data Analysis

First, the word repetitions were parsed from the carrier phrase using SMASH (Green et al., 2013). 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/.

Then, a custom-written MATLAB script was used to analyze the vertical movements of the jaw and lower lip during the production of "muffin". Lower lip 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 positional minimum for the open vowel /ʌ/ and the positional maximum for the labiodental fricative /f/. Although this approach differs from the traditional phase angle calculations, it is well-suited to quantify the inter-articulatory timing patterns of the lower lip and jaw.

For better spatial alignment and visual inspection of the kinematic data, 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 (lag) because the study sought to determine the strength of lower lip and jaw coupling. In other words, as a first step, we merely investigated differences in the absolute lag times between the jaw and the lower lips. The order in which the lip and jaw reached the target was not of interest at this point. Because lag times may be more difficult to interpret when talkers 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 trial-to-trial variability in the lag times across five repetitions, we also calculated the coefficient of variation (CoV) based on the

absolute lag values. The CoV was defined as the standard deviation across five repetitions divided by the talker's mean lag time across five repetitions.

2.4. Statistical Analysis

Linear mixed models were completed to determine between-group differences in absolute and percent lag times with group 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 between-group differences. Because of the preliminary nature of the study, a critical alpha-level of $p < .05$ was selected for all test and Cohen's d effect sizes were calculated. Absolute Cohen's d < .4 and < .8 were interpreted as small and medium effects, respectively. A negative Cohen's d indicated that Group 1's mean was smaller than Group 2's mean in comparison.

3. Results

Group means (*SE*) of each dependent variable are provided in **Table 1**. The group means for the duration of the utterance "muffin" are also shown to better interpret the absolute lag durations and the %lag durations. No significant between-group differences were found for the absolute and the percent lag times as well as for the CoV of the lags for either target. Nevertheless, as can be seen in **Table 2,** medium to large effect sizes were observed for absolute lag times of both targets for ALS vs. controls. Furthermore, medium effect sizes were found for ALS vs. PD. That is, for both targets, talkers with ALS tended to have longer absolute lag times than talkers with PD and/or controls. Effect sizes for comparisons between talkers with PD and controls were small for both targets.

When considering %lag, the medium to large effects for comparisons between talkers with ALS and controls went away for both targets. Medium effects for ALS vs. PD turned in the opposite direction and only remained at a medium size for the target /f/, but diminished to a small effect for the target /ʌ/. In addition, the small effect sizes for PD vs. controls increased slightly from a small to a medium effect for /f/ while the small effect for /ʌ/ went away almost completely.

For both targets, trial-to-trial variability in lag times (CoV) tended to be lower in both clinical groups relative to those of controls. Furthermore, talkers with ALS had lower CoV values than talkers with PD. In fact, large and medium differences were observed between talkers with ALS and PD for the target /ʌ/ and /f/, respectively, while medium and small differences were observed between talkers with ALS and controls, respectively. By contrast, small differences in CoV were observed between talkers with PD and controls for both targets.

3.1. Figures and Tables

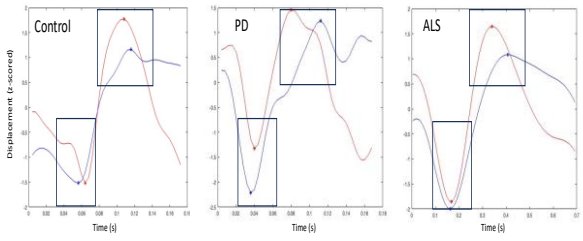


Figure 1: *Spatially normalized (z-scored) movements of the lower lip (red) and jaw (blue) during the word "muffin". Shaded areas highlight the troughs/ peaks for targets /ʌ/ and /f/, respectively.*

Table 1: *Group means (SE) for all dependent variables*.

Group	Control	PD	ALS
Lag_/ʌ/	.004 (.001)	.005 (.001)	.010 (.003)
%Lag /ʌ/	2.449 (.471)	2.730 (.560)	2.344 (.394)
CoV Lag /ʌ/	.737 (.178)	.599 (.052)	.306 (.105)
Lag_/f/	.011 (.002)	.016 (.004)	.027 (.011)
%Lag /f/	6.278 (.831)	9.049 (2.696)	5.449 (1.430)
CoV Lag /f/	.963 (.294)	.852 (.107)	.611 (.161)
Total Word Duration	.185 (.012)	.190 (.014)	.479 (.140)

Table 2: *Effect sizes (Cohen's d) for group comparisons*. Large effect sizes indicated in **bold**

Variable	ALS vs. Controls	PD vs. Controls	ALS vs. PD
Lag_/ʌ/	**1.02**	.21	.62
%Lag /ʌ/	-.06	.01	-.25
CoV Lag /ʌ/	-.73	-.24	**-1.05**
Lag_/f/	.78	.35	.40
%Lag /f/	-.23	.40	-.54
CoV Lag /f/	-.36	-.11	-.52

4. Discussion and Conclusion

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 interarticulator coupling was consistent across five repetitions of the same utterance (i.e., CoV) and compared these findings across two clinical groups with different underlying impairments of the speech motor system (PD and ALS) relative to healthy controls. It was hypothesized that talkers with ALS would exhibit stronger interarticulator 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 the primary articulators (i.e., tongue, lower lip). Preliminary findings for lag times did not support this hypothesis because talkers with ALS tended to have longer absolute lag times than controls and talkers with PD for both targets. Therefore, their jaw and lower lip movements appear to be less synchronized than those of the other two groups. However, there was a trend toward shorter relative lag times (%lag) in talkers with ALS when compared to the other talkers. Future studies should determine the extent speech rate differences impact lag times. Such insights would help the interpretation of the findings for talkers with ALS in this study.

With regards to trial-to-trial variability, talkers with ALS tended to show more consistent lag times than controls and talkers with PD, regardless of the target. This finding aligns with previous work showing lower spatiotemporal pattern variability in talkers with ALS (e.g., Kuruvilla-Dugdale & Mefferd, 2017) and may suggest that these talkers have less flexibility in their articulatory system to change lower lip-jaw coupling. However, given the preliminary nature due to the small sample size, further research is warranted to replicate this finding.

Given the basal ganglia pathology, we also expected deviant lag times for talkers with PD. However, effect sizes for comparisons between talkers with PD and controls were in general small and suggested only minimal differences in the

lower lip-jaw interarticulatory timing patterns across these two talker groups, particularly for the target /ʌ/. This finding may provide support for the notion that articulatory movements of talkers with PD are generally merely downscaled in size while interarticulatory timing patterns are being preserved. The findings of the current study are preliminary and need to be replicated; however, they suggest that although basal ganglia pathologies can disrupt absolute temporal timing patterns (e.g., segment durations, speech rate), they may not disrupt relative timing such as lip-jaw interarticulatory timing patterns.

The trial-to-trial variability of the lag values were rather comparable between talkers with PD and controls considering the small effect sizes between these two groups. Thus, talkers with PD may have an unaffected flexibility to modify their interarticulator coupling. Larger sample sizes are needed to solidify the observed trends of the current study.

In sum, the study findings support our current conceptual understanding of timing disruptions in talkers with impaired motor speech systems. That is, this study provides preliminary evidence of a mismatch between the time designated by the central clock and the physical state of the articulators in talkers with ALS. Furthermore, our findings suggest that despite the basal ganglia pathology, the relative timing patterns among articulators appears rather intact in talkers with PD. Future studies should examine such aspects of articulatory timing behavior in talkers with other basal ganglia pathologies (i.e., Huntington's disease) as well as contrast current findings with those of talkers with cerebellar pathologies. Such studies will help to better understand the disease-specific pathomechanisms affecting articulatory timing in talkers with dysarthria.

5. Acknowledgements

This research was funded by NIH-NIDCD grants 1R15DC016383 and R21DC019952-01 (PI: Kuruvilla-Dugdale), and grants R03DC015075 and R01DC019648-01A1 (PI: Mefferd). We are grateful to the research assistants and subjects who participated in the study. Special thanks to Emily Beultel, Emmersen Haugland, Thushani Munasinghe, Chaewon Park, and Olivia Stanislawski for helping with data analysis.

6. References

Balci, F., Ludvig, E. A., Abner, R., Zhuang, X., Poon, P., & Brunner, D. (2010). Motivational effects on interval timing in dopamine transporter (DAT) knockdown mice. *Brain research*, *1325*, 89-99.

DePaul, R., Abbs, J. H., Caligiuri, M., Gracco, V. L., & Brooks, B. R. (1988). Hypoglossal, trigeminal, and facial motoneuron involvement in amyotrophic lateral sclerosis. *Neurology*, *38*(2), 281-281.

Gracco, V. L. (1988). Timing factors in the coordination of speech movements. *Journal of Neuroscience*, *8*(12), 4628-4639.

Grahn, J. A. (2009). The role of the basal ganglia in beat perception: neuroimaging and neuropsychological investigations. *Annals of the New York Academy of Sciences*, *1169*(1), 35-45.

Green, J. R., Wang, J., & Wilson, D. L. (2013, September). SMASH: a tool for articulatory data processing and analysis. In *Interspeech* (pp. 1331-1335).

Harrington, D. L., Haaland, K. Y., & Hermanowitz, N. (1998). Temporal processing in the basal ganglia. *Neuropsychology*, *12*(1), 3.

Hermes, A., Mücke, D., Thies, T., & Barbe, M. T. (2019). Coordination patterns in Essential Tremor patients with Deep Brain Stimulation: Syllables with low and high complexity. *Laboratory Phonology*, *10*(1).

Konoike, N., Kotozaki, Y., Miyachi, S., Miyauchi, C. M., Yomogida, Y., Akimoto, Y., ... & Nakamura, K. (2012). Rhythm information represented in the fronto-parieto-cerebellar motor system. *Neuroimage*, *63*(1), 328-338.

Kuruvilla-Dugdale, M., & Mefferd, A. (2017). Spatiotemporal movement variability in ALS: Speaking rate effects on tongue, lower lip, and jaw motor control. *Journal of communication disorders*, *67*, 22-34.

Lake, J. I., & Meck, W. H. (2013). Differential effects of amphetamine and haloperidol on temporal reproduction: dopaminergic regulation of attention and clock speed. *Neuropsychologia*, *51*(2), 284-292.

Langmore, S. E., & Lehman, M. E. (1994). Physiologic deficits in the orofacial system underlying dysarthria in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, *37*(1), 28-37.

Lee, J., & Bell, M. (2018). Articulatory range of movement in individuals with dysarthria secondary to amyotrophic lateral sclerosis. *American Journal of Speech-Language Pathology*, *27*(3), 996-1009.

MacDonald, C. J., & Meck, W. H. (2005). Differential effects of clozapine and haloperidol on interval timing in the supraseconds range. *Psychopharmacology*, *182*, 232-244.

Malapani, C., Rakitin, B., Levy, R., Meck, W. H., Deweer, B., Dubois, B., & Gibbon, J. (1998). Coupled temporal memories in Parkinson's disease: a dopamine-related dysfunction. *Journal of cognitive neuroscience*, *10*(3), 316-331.

Maricq, A. V., & Church, R. M. (1983). The differential effects of haloperidol and methamphetamine on time estimation in the rat. *Psychopharmacology*, *79*, 10-15.

Meck, W. H. (2005). Neuropsychology of timing and time perception. *Brain and cognition*, *58*(1), 1-8.

Mefferd, A. S. (2017). Tongue-and jaw-specific contributions to acoustic vowel contrast changes in the diphthong/ai/in response to slow, loud, and clear speech. *Journal of Speech, Language, and Hearing Research*, *60*(11), 3144-3158.

Mooshammer, C., Hoole, P., & Geumann, A. (2006). Interarticulator cohesion within coronal consonant production. *The Journal of the Acoustical Society of America*, *120*(2), 1028-1039.

Rong, P., & Heidrick, L. (2022). Functional Role of Temporal Patterning of Articulation in Speech Production: A Novel Perspective Toward Global Timing–Based Motor Speech Assessment and Rehabilitation. *Journal of Speech, Language, and Hearing Research*, *65*(12), 4577-4607.

Saltzman, E., Nam, H., Krivokapic, J., & Goldstein, L. (2008, May). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proceedings of the 4th international conference on speech prosody (speech prosody 2008), Campinas, Brazil* (pp. 175-184).

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4), 333-382.

Shellikeri, S., Green, J. R., Kulkarni, M., Rong, P., Martino, R., Zinman, L., & Yunusova, Y. (2016). Speech movement measures as markers of bulbar disease in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, *59*(5), 887-899.

Tanaka, S. C., Schweighofer, N., Asahi, S., Shishida, K., Okamoto, Y., Yamawaki, S., & Doya, K. (2007). Serotonin differentially regulates short-and long-term prediction of rewards in the ventral and dorsal striatum. *PloS one*, *2*(12), e1333.

Windmann, A., Šimko, J., & Wagner, P. (2015). Optimization-based modeling of speech timing. *Speech Communication*, *74*, 76-92.

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, USA*
[2]*Adelphi University, USA*
[3]*Haskins Laboratories, USA*

hitchcocke@montclair.edu, lkoenig@adelphi.edu, laura.koenig@yale.edu

Abstract

Do children with speech-sound disorders (SSDs) also differ in their speech perception? Past results suggest that perceptual difficulties are limited to sounds produced in error. Here, we assessed labeling accuracy and reaction times [RTs] in children with SSD (without voicing errors) and typical-developing [TD] peers. Stimuli were words 'boo, Pooh, doe, toe' produced by TD 2-year-olds, with VOTs that were "appropriate" (expected for the target) or inappropriate. Listener judgments were considered accurate if they matched the child's target. Results showed high listener accuracy for appropriate VOTs with no group differences. For inappropriate VOTs, children with SSD showed higher accuracy than TD, reaching significance for one comparison. RTs were faster for accurate labeling in both groups and were overall shorter children with SSD than TD peers, suggesting that children with SSD may demonstrate some differences in speech perception behavior, even for sounds not produced in error.

Keywords: speech perception, speech sound disorders, reaction time

1. 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, finding differences in 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 produced by young children. As in past work, we present data on labeling accuracy; we also add a preliminary analysis of reaction times [RTs].

2. Methods and Analysis

2.1. Participants

Listening participants included 15 monolingual English-speaking typically-developing children (TD: 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 function, 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.

2.2. Listening task and stimuli

Listeners were asked to perform a forced-choice identification task in response to child-produced 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", "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/ values. 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 (see **Figure 1**). 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**). Each listener provided 288 responses (4 target words x 6 child speakers x 6 exemplars per child speaker x 2 VOT categories, viz. appropriate and inappropriate for the target), yielding 8352 datapoints.

The design of the stimulus set (viz., toddler-produced words chosen to have bimodal VOT distributions) is a continuation from previous work (Hitchcock & Koenig, 2021). In the current context, we note the following: a) The variability inherent in young children's speech may increase the level of difficulty for listeners, i.e. provide a more sensitive test of group differences. b) Conversely, the separation between target /b d/ and /p t/

within the short- and long-lag VOT regions may aid listeners in ascertaining the child's target.

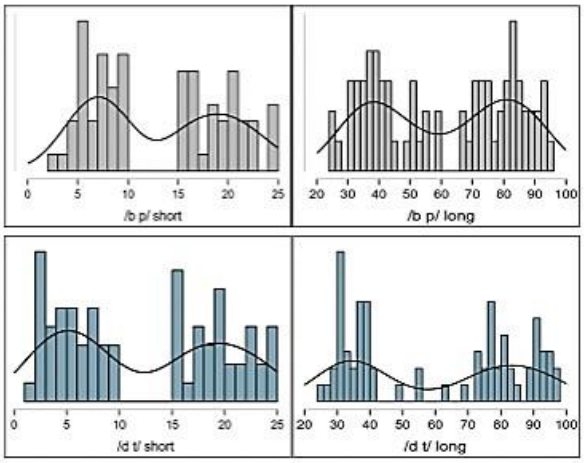


Figure 1: *Distribution of stimuli along the VOT continuum.*

2.3. Measures and processing

2.3.1 Accuracy

We classified whether listener ratings (phoneme labels) were accurate (defined as matching the speaker's intended target), and assessed RTs as described in the next paragraph. Note that children were not instructed to respond as quickly as possible.

2.3.2 Response times

Since the RTs were positively skewed, we first log-transformed the data (values that we henceforth call logRTs). We removed original RTs that were negative, which could not be log-transformed and presumably represented false starts. This removed 156 tokens from the dataset, with tokens heavily concentrated in SSD children (148/156 =95%). Three children with SSD accounted for 119 of these values. In the most extreme case (58 removed cases), we still had 83% of the child's data to analyze. We then z-transformed the data (based on the mean and SD of the full dataset), yielding logRTz. Finally, we removed logRTz values that were > |3| standard deviations from the grand mean. The final trimmed logRTz dataset contained 8084 productions.

3. Results

3.1. Accuracy

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: long-lag productions of /b d/ and short-lag productions of /p t/. Results are presented in **Figures 2–3** and statistical results are summarized 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 within VOT categories. Group differences were only significant for one comparison (long-lag/inappropriate /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). This can be seen in **Figures 2–3**.

Table 1: *Statistics on group differences (Mann-Whitney U-values and associated p-values) for all stop consonants, with appropriate and inappropriate VOT values.*

	b	d	p	t
Appropriate VOTs				
U value	133200	135468	133650	135288
p-value	0.279	0.824	0.124	0.574
Inappropriate VOTs				
U value	128016	135792	129960	128916
p-value	*0.046	0.943	0.119	0.083

*Indicates statistical significance (p <.05).

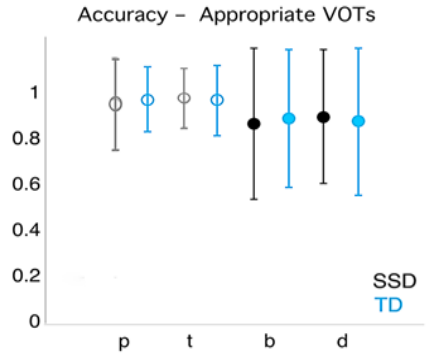


Figure 2: *Accuracy means and standard deviations for both groups – Appropriate VOT values.*

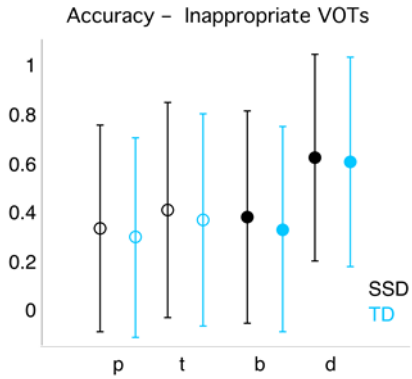


Figure 3: *Accuracy means and standard deviations for both groups – Inappropriate VOT values.*

In all cases, accuracy was much higher for appropriate VOTs (**Figure 2**) than for inappropriate VOTs (**Figure 3**) suggesting that listener judgments were mainly driven mainly by VOT. Variability is extensive in both SSD and TD groups. Unexpected high accuracy in both groups for one inappropriate VOT condition (long-lag /d/, **Figure 3**) could reflect secondary cues available in the stimuli.

3.2. Reaction times

Levene's tests of variance equality were significant across groups and accuracy measures, so we employed non-parametric statistics to test for group differences.

The data show shorter RTs for the SSD group than the TD group (SSD mean = -0.061, SD = 0.176; TD mean = 0.030, SD =

0.184). We also find shorter RTs for accurate responses than inaccurate (Accurate mean = -0.030, SD = 0.177; Inaccurate mean = 0.024, SD = 0.187). Data, split by group and accurate/inaccurate responses, are shown in **Figure 4**.

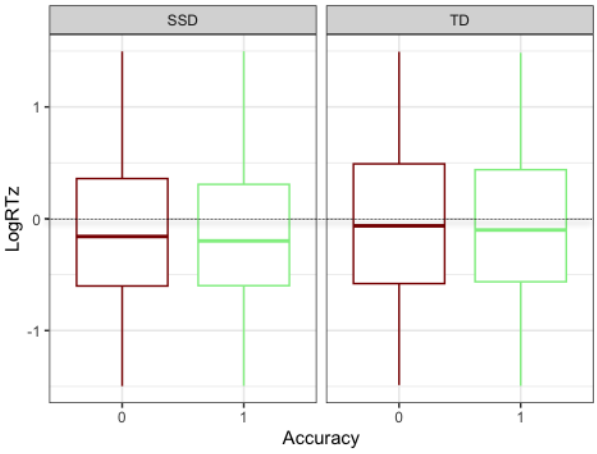


Figure 4: *Logged and z-transformed reaction times as a function of group and response accuracy (0=inaccurate, 1=accurate). The horizontal line at zero is intended to facilitate group comparisons. Outliers have been trimmed from the display.*

For both inaccurate and accurate responses, Kruskal-Wallis tests showed a highly significant group difference in logRTz (SSD < TD): Accurate responses, $\chi^2 = 14.400$, df = 1, p<0.001; inaccurate responses, $\chi^2 = 10.691$, df = 1, p-value = 0.001.

Evaluating whether logRTz values differed within groups as a function of accurate and inaccurate responses, we find a significant difference in the TD group: $\chi^2 = 4.002$, df = 1, p-value = 0.046. This did not hold for the SSD group: $\chi^2 = 1.032$, df = 1, p-value = 0.310.

Finally, we asked whether response speed differed depending on whether VOTs were appropriate or inappropriate for the target. Again, we observe a significant difference in the TD group ($\chi^2 = 5.359$, df = 1, p-value = 0.021) but not the SSD group: $\chi^2 = 0.783$, df = 1, p-value = 0.376. For both appropriate and inappropriate VOTs, the group difference (SSD < TD) remained significant.

As a precaution, we removed the three SSD children who contributed the greatest number of false starts and re-evaluated these conclusions (reduced dataset containing 3106 and 4261 datapoints for SSD and TD groups, respectively). Group differences remained significant in all cases. Median values are provided in Table 2. As seen before, a) all values are lower for SSD than TD; b) accurate responses are lower (faster) than inaccurate, and c) responses to appropriate VOTs are faster than to inappropriate VOTs. These values demonstrate (see also **Figures 2–3**) that group differences are quite modest.

Finally, **Figure 5** shows the logRTz values for individual listeners in both groups. There is considerable group overlap at the low end (faster RTs), but the groups diverge at the high end (slower RTs).

Table 2: Median LogRTz values divided by group (SSD, TD), response accuracy (inaccurate, accurate), and target VOT (inappropriate, appropriate).

	Response		Stimulus VOT	
	Inacc.	Acc.	Inapp.	App.
SSD	-0.115	-0.161	-0.136	-0.162
TD	0.001	-0.076	-0.005	-0.083

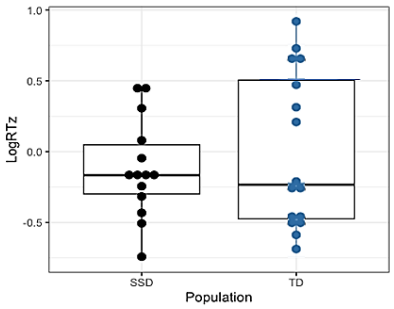


Figure 5: *Median LogRTz values for all speakers in both groups.*

4. Discussion and conclusion

4.1. Accuracy

In both child groups, labeling was highly accurate for targets with appropriate VOTs. This is consistent with previous work showing high accuracy in adults for young child productions with appropriate VOT values (Hitchcock & Koenig, 2021). The statistical results for perceptual accuracy are also generally 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 *in*appropriate VOT targets in children with SSD warrants further investigation and could suggest subtle perceptual differences between groups that are not seen in other testing paradigms. Potentially, children with SSD could have less refined 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. The adult responses to inappropriate VOT values for /p t/ were considerably lower than those 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. To the extent that distributional characteristics of the data influenced listener responses in the current work, it appears to have had largely similar effects in both TD and SSD groups.

4.2. Reaction times

Reaction time (logRTz) data were slower for inappropriate VOTs in both groups, as one might expect. LogRTz's were also slower for inaccurate responses in both groups. Importantly, this held for both groups, and moreover, rating accuracy did not differ greatly between groups. Follow-up analyses could assess only correct responses, but this would lead to high data loss in some of the VOT categories and limit sensitivity to group differences. Perhaps the most surprising finding is that the SSD

group had faster reaction times, and this difference remained significant regardless of accurate/inaccurate responses, appropriate/ inappropriate VOTs, and removing children who had atypical (false-start) RTs. This finding, though tentative, deserves further exploration and could indicate some differences between how SSD and TD children process speech, or respond to a task like the one we presented here.

4.3. General conclusions

Listener accuracy for SSD and TD groups was largely comparable, in line with past work suggesting that children with SSD do not show clear speech perception difficulties for non-errored sounds. Interestingly, however, for inappropriate VOTs the SSD group, on average, tended to out-perform their TD peers, and this was significant in one of four comparisons. This result deserves further exploration. As part of this, we will explore individual differences among the listeners (Kong & Edwards, 2016). We also plan to assess how secondary cues in the stimuli (durational measures, f0, burst intensity) might have contributed to listener responses in TD and SSD groups.

The current RT results speak against the notion that children with SSD have a general speech perception difficulty that is manifested in slower responses, at least for non-errored sounds. Nevertheless, this modest dataset does not allow us to assert with confidence that children with speech-sound disorders are universally faster in their phonetic labeling. Along with replicating these results in larger listener groups, future work should employ more sophisticated modeling to tease apart the many possible inter-relationships among group, VOT category, response accuracy, etc.

5. Acknowledgements

The authors would like to thank the participants and their families for their ongoing cooperation throughout the study. We also express our thanks to graduate research assistants Madeline Cheyne, Amy Rosen and Dana Catalano, graduate research assistants for their support with data collection and management.

6. 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.

Kong, E. J., & Edwards, J. (2016). Individual differences in categorica l perception of speech: Cue weighting and executive function. *Jou rnal of Phonetics*, 59, 40-57.

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 discrim-ination/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.