Proceedings of the Nineteenth

Australasian International Conference on Speech Science and Technology

3–5 December 2024, Melbourne, Australia

Welcome
THE UNIVERSITY OF MELBOURNE
Wominjeka
The University of Melbourne acknowledges the Aboriginal and Torres Strait Islander traditional owners of the unceded land on which we work and learn. We pay respect to the Elders, past and present, and the place of Indigenous knowledge in the academy.

Editors: Olga Maxwell & Rikke Bundgaard-Nielsen

Cover design: Susie Nahm

Conference Information

Table of Contents

Author Index

Search

Proceedings of the Nineteenth Australasian International Conference on Speech Science and Technology. ISSN 2207-1296.

ASSTA
The Australasian Speech Science & Technology Association Incorporated

Acknowledgement of Country

As we welcome delegates of the 19th International Speech Science and Technology (SST) Conference from across Australia, New Zealand and the world to The University of Melbourne, we recognise the Traditional Owners of the unceded land on which we work, learn and live: the Wurundjeri Woi-wurrung and Bunurong peoples (Burnley, Fishermans Bend, Parkville, Southbank and Werribee campuses), the Yorta Yorta Nation (Dookie and Shepparton campuses), and the Dja Dja Wurrung people (Creswick campus). The biannual SST Conferences are multidisciplinary celebrations of rich linguistic diversity and complexity with substantial impact across a wide range of disciplines. We are conscious, however, that the world's linguistic diversity remains under threat, and particularly conscious of the language loss experienced by many Indigenous Peoples in the world, including the Wurundjeri Woi-wurrung on whose lands we host this event. We are similarly conscious that many of the World's languages, including many Australian languages, are yet to be afforded detailed study. As a research community, we recognise our obligation to respond to this and extend our academic focus to these languages and our support to these language communities.

Corroboree. Drawing by Kwat Kwat artist Yakaduna [Tommy McRae], c1890, featuring a line of dancing Wathaurong men, of the Kulin Nation, with a white-skinned man—William Buckley—separate from, but aligned with the dancers, but for his wide brimmed hat. The central Arts West building at The University of Melbourne incorporates Yakaduna's *Corroboree* drawing in its sculptured metal façade.

Welcome to Delegates

On behalf of the organising committee and the Australasian Speech Science and Technology Association (ASSTA), we warmly welcome you to the 19th Australasian International Conference on Speech Science and Technology (SST2024). It is wonderful to return to Melbourne after 14 years, this time hosted by The University of Melbourne on our Parkville campus, close to the historic city centre.

As one of Australia's key hubs for speech science and technology, Melbourne is the ideal location for this international event, which brings together researchers across disciplines, including phonetics, engineering, forensics, psycholinguistics, audiology and speech pathology, and linguistics. This year, we are pleased to welcome delegates from Australia, New Zealand, UK, France, Japan, Malaysia, Switzerland, Germany, Belgium, China, India, Canada, Italy, and Sweden. We also have a strong representation from graduate scholars and early career researchers. Our conference program covers a broad range of topics and includes 62 high-quality submissions, delivered as 46 oral presentations and 16 posters, with 54 papers forming part of the conference proceedings. Altogether, there will be many opportunities to share research, explore methodological advances, and engage in theoretical discussions.

We are pleased to welcome organisers and presenters of the two special sessions: *The Phonetic Expression of Phonological Length* and *Sociophonetic Variability in Australian English Varieties (SocioPhonAus4)*, with the latter representing the growing community of researchers working on varieties of English in Australia and New Zealand.

We are especially thrilled to host four distinguished female keynote speakers this year: Marija Tabain, Sasha Calhoun, Rosey Billington, and Lynn Clark. Marija Tabain will open the conference with a presentation on proposed revisions to the current IPA consonant chart, while Rosey Billington will give the inaugural Anne Cutler Lecture, honouring Professor Cutler's significant contribution to speech sciences. Keynotes Sasha Calhoun and Lynn Clark, who joined us from New Zealand, will deliver keynotes on the final day, with Lynn Clark presenting as part of the *SocioPhonAus4* special session. We also welcome Rosey Billington, Sasha Calhoun, Brett Swanson, Joshua Penny, and Hannah White as tutorial presenters.

We would like to acknowledge the University of Melbourne for its support of the conference, in particular the School of Languages and Linguistics, the Department of Biomedical Engineering and the Faculty of Arts, who have provided financial support. We are extremely grateful to the School of Language and Linguistics for the invaluable support with conference planning and organisation.

We extend a warm thank you to all volunteer reviewers, authors, and delegates, whose contributions and enthusiasm enrich this conference. We hope you find plenty of stimulating discussions and networking opportunities at the conference, and that this meeting leads to new collaborations and research projects. Enjoy SST2024!

Olga Maxwell,

Conference Chair

Conference Hosts

ASSTA

The Australasian Speech
Science & Technology
Association Incorporated

THE UNIVERSITY OF
MELBOURNE

School of Languages and Linguistics

Department of Biomedical Engineering

Faculty of Arts

ASSTA

The Australasian Speech
Science & Technology
Association Incorporated

2024 ASSTA New Researcher Award

Elise Tobin
Macquarie University

Conference Organisation

Organising Committee

Olga Maxwell (Chair)
Debbie Loakes (Secretary)
David Grayden (Treasurer)
Janet Fletcher
Chloé Diskin-Holdaway
John Hajek
Yizhou Wang (Publicity)
Angelo Dian (Student Liaison)

Programme Committee

Olga Maxwell
Rikke Bundgaard-Nielsen
David Deterding
Helen Fraser
Brett Baker
Colleen Holt
James Walker

Conference Support

Shuting Liu
Lourdes Sanchez Guerrero
Anna Henger

Volunteers

Canaan Lan
Josh Clothier
Rana Hosseinpoor Damirchian
Michael Sadeghi
Fatemeh Aghaei
Geordie Kidd
Dan Weber
Yiran Guo

Review Panel

Mark Antoniou
Brett Baker
Kirrie Ballard
Rosey Billington
Jason Brown
Laurence Bruggeman
Francesco Burroni
Denis Burnham
Sasha Calhoun
Josh Clothier
Felicity Cox
Katherine Demuth
Angelo Dian
Gerry Docherty
Julien Epps
Paola Escudero
Lois Fairclough
Paul Foulkes

Ksenia Gnevsheva
Simón Gonzalez
James Grama
David Grayden
Adele Gregory
John Hajek
Hilmi Hamzah
Rebecca Holt
Vincent Hughes
Justine Hui
Jesin James
Kathleen Jepson
Craig Jin
Ghada Khattab
Jeesun Kim
Yuko Kinoshita
Liquan Liu
Debbie Loakes

Kirsty McDougall
Joshua Penney
Michael Proctor
Gan Qiao
Michael Robb
Phil Rose
Vidhyasaharan Sethu
Elena Sheard
Chilin Shih
Steven So
Tünde Szalay
Marija Tabain
Catherine Theys
Catalina Torres
Yizhou Wang
Catherine Watson
Hannah White
Ivan Yuen

Nineteenth Australasian International Conference on Speech Science and Technology

SST 2024

Table of Contents

Day 1: 3 December 2024

Keynote

Session 1: Consonants

Session 2: Machine Learning

Session 3: Speech Processing

Session 4: Vowels

Session 5: Bilingual Speech

Session 6: Australian English

Day 2: 4 December 2024

Session 7: Child Language Acquisition

Special Session: The Phonetic Expression of Phonological Length

Session 8: Voice Quality and Prosody

Poster Session

Day 3: 5 December 2024

Special Session: Sociophonetic Variability in the English Varieties of Australia

Session 9: Forensics

Session 10: Affective Speech

A Consideration of the Consonant Chart in the International Phonetic Alphabet

Marija Tabain[1], *Marc Garellek*[2], *Matthew Gordon*[3]

[1]La Trobe University, [2] University of California San Diego, [3] University of California Santa Barbara

m.tabain@latrobe.edu.au, mgarellek@ucsd.edu, mgordon@linguistics.ucsb.edu

Abstract

In this paper we outline various problems with the current IPA consonant chart, based on the study of an ever-increasing number of languages. We propose a revised version of the chart, which has many similarities with the pre-1989 version of the chart, but which at the same time shows innovations. We particularly focus on the laryngeal sounds; the epiglottal and pharyngeal sounds; the retroflexes; the sibilant fricatives; and the semi-vowels. We hope that our discussion will be useful to both researchers and pedagogues, and we hope that our proposed revisions are logical from the phonetic and phonological points of view.

Index Terms: IPA chart, consonants, revision

1. Introduction

Human speech sounds are complex, with some sounds (such as vowels) existing more in a gradient space, and some sounds (such as some consonant contrasts) being more categorical. Stops and fricatives might be said to be more categorical, in that if a speech sound involves full blockage of airflow, it is a stop; and if the sound involves sustained noise shaped by the oral cavity, it is a fricative. There cannot be said to be an intermediate state where the sound is between a stop and a fricative in acoustic or articulatory terms. At the same time, one could argue that some contrasts of place are more gradient, since one could describe a sound as being "more velar" or "more uvular", or "more palatal".

Given such complexity in speech sounds, it is perhaps not surprising that over the decades, the chart of the International Phonetic Alphabet (IPA) has been revised several times, as phoneticians' understanding of speech is improved, partly through the use of instrumental techniques, and partly through the study of a greater variety of languages – historical IPA charts are given at https://www.internationalphoneticassociation.org/content/ipa-chart. We assume that the reader is familiar with the most recent chart of the IPA. However, throughout this paper we will refer to the 1951 chart of the IPA (which is also available at the above website), since this was the most recent chart before the significant reforms of the Kiel Convention in 1989 [1].

In this paper we consider some problems with the current IPA consonant chart, based on current knowledge of articulations in a variety of languages, and based on our own classroom experience of the most confusing aspects of the chart when teaching these sounds. We offer a possible alternative solution to the consonant chart, and while we recognize that our suggestions are not perfect, we hope that it is conceptually clearer and more accurate than the current chart, and that it will stimulate further discussion around the chart.

2. Some issues with place of articulation

We begin by considering some problems with the consonant chart in terms of place of articulation:

2.1. Glottal

The Glottal column contains only a voiceless "plosive" and two "fricatives", with every other cell greyed out because the articulations are impossible (e.g. a glottal lateral). The existing symbols are problematic in terms of their voicing, in that the glottal stop [ʔ] almost never involves full closure (and nothing that could really be termed a "burst", as other plosives almost always have). Therefore, it may not be appropriate to place [ʔ] with other "voiceless" consonants. Glottal stop also does not pattern with other voiceless plosives in blocking the spread of nasality or vowel features in phonological harmony. The characteristic lack of a full closure is consistent with the frequent patterning of [ʔ] as a prosodic feature rather than a consonant phonologically, e.g. in weight-sensitive stress systems [2]. A recent survey found a similar phonetic realization for glottal "fricatives" [h ɦ], whereby the voicing for both sounds is largely predictable by context: the voiced glottal fricative is more common intervocalically, and the voiceless more common at word edges. Although this conditioning environment can be said to apply for other stop and fricative places of articulation (e.g. [p b]), the glottals are different in that contrasts between /h/ and /ɦ/ are controversial and likely reflect phonetic and phonological features other than voicing [3]. This is in stark contrast to the many languages that are known to have contrasts in stop or fricative voicing.

At the same time, the glottal "fricatives" are problematic as fricatives, if we consider frication to be turbulent noise that is generated by a constriction in the vocal tract. For glottal "fricatives", noise is generated when the vocal folds are held apart – this is closer to the definition of an approximant. In addition, acoustically the intensity of the noise is not comparable to the intensity of the oral fricatives. Indeed, the glottal /h/ is often described as the voiceless version of the adjacent vowel. Overall, the glottal fricatives are not easily defined from an articulatory or even an acoustic point of view. Instead, it is perhaps more appropriate to say that [h ɦ] is glottal spreading, and [ʔ] (which is very often not a stop) is glottal constriction [3, 4].

Studies involving direct imaging of [ʔ] have also shown that glottal constriction is usually produced with supraglottal laryngeal constriction, especially of the ventricular folds but also of general epilaryngeal constriction. If this supraglottal constriction is deemed criterial for producing a glottal stop, then perhaps it should be relabeled "laryngeal" instead of "glottal" [5]. This would be a return to a label used in the early days of the IPA (e.g. in 1900, 1903, 1905, and 1921) [6] and would find support in phonological feature geometric classifications that place the glottal sounds under a "Laryngeal" node [7, 8].

Perhaps more importantly, glottal consonants pattern differently from other consonants in many phonological processes across the world's languages. For instance, in Sundanese, a process spreading nasalization across the word is blocked by a non-nasal consonant – however, glottals are transparent for this purpose (i.e. they do not behave as a consonant in this regard) [9, 10]. Similarly, glottals characteristically stand out from other consonants in allowing complete harmony (i.e. vowel copying) between vowels in adjacent syllables, e.g. in Mesoamerican languages [11, 12]. Glottals also behave as transparent segments, unlike other consonants, in child language phonology [13]. Even in English, the allomorphic rule assigning the indefinite article to a noun phrase treats the /h/ as transparent (i.e. as a non-consonant) for some speakers – e.g. /ən hɪstɔɹɪk əvent/. All of these observations are in line with the above-mentioned feature geometry approach to phonology, where glottal sounds are located within their own node "Laryngeal", quite separate from the node "Place".

For all of the above reasons, we suggest that the glottal column be labelled "Laryngeal". This is in line with the Laryngeal Articulator Model [5] that treats glottal stop as involving multiple laryngeal structures. However, in order to respect the unusual status of these sounds with respect to all of the other consonants in the chart, we add a thicker line between the laryngeal column and the rest of the consonant chart, in order to highlight the fact that although laryngeals often pattern as consonants, there are many occasions where they are transparent to phonological processes that apply to (all) other consonants. In addition, we have marked cell boundaries with dotted rather than solid lines, in order to denote that the standard supralaryngeal manner definitions are not relevant for laryngeals. At the same time, we place [h ɦ] in the approximant row, to denote that the articulatory gesture involves spreading. There is thus a visual distance between [ʔ] and [h ɦ] which is indicative of the continuum of laryngeal constriction.

In line with proposals and subsequent discussion following the Kiel Convention [4, 14], we have considered simply removing the laryngeals from the consonant chart altogether, and placing them on their own separate line outside of the chart – this choice would particularly highlight the problems with the "stop" and "fricative" manners of articulation. However, we felt that this choice would be more problematic in terms of backward compatibility of the chart, and would also deny the laryngeals a status of (albeit imperfect) consonants. For this reason we choose the thicker line as the more backward compatible option.

Our draft revision of the IPA chart is included in the Appendix to this paper.

2.2. Pharyngeal and Epilaryngeal

In the present IPA chart, the column "Pharyngeal" has only two symbols listed, namely the voiced and voiceless fricatives [ħ ʕ]. However, the voiced pharyngeal [ʕ] is more often realized as an approximant rather than a fricative. Indeed, the chart allows the possibility of a pharyngeal approximant, as well as a voiceless pharyngeal plosive, and also a trill, tap or flap.

At the same time, the voiceless and voiced epiglottal fricatives [ʜ ʢ], and the epiglottal plosive [ʡ] are listed under "Other symbols" beneath the main consonant chart. It should be noted that the epiglottal plosive [ʡ] has been characterized as a pharyngealized glottal stop [5].

We propose to merge the categories pharyngeal and epiglottal into a single column, labelled "Pharyngeal and Epilaryngeal" (see Appendix). Note that we write "Epilaryngeal" instead of "Epiglottal", to highlight the role that the superior larynx plays in producing lower-pharyngeal/epiglottal constriction [5]. The voiceless stop in this column is [ʡ], with no voiced stop counterpart, in line with the view that it is an epilaryngeal (or ‘pharyngealized glottal’) stop. We propose two trills for this column, voiceless [ʜ] and voiced [ʢ], in line with the typical realization of these two epiglottals (see for example Figure 5.23 on page 168 of [15] which shows clear trilling for [ʜ] in the North Caucasian language Agul). We note that this will be the first voiceless member of the trill manner row (though they occur quite often phonetically as variants of the voiced trills). The voiceless pharyngeal fricative [ħ] occupies the fricative cell in this column, and the voiced pharyngeal [ʕ] occupies the approximant cell in this column.

2.2.1. A short note on the uvular and velar places of articulation

We recognize that the uvular fricatives, like pharyngeals, are also often not realized as fricatives: the voiceless [χ] is often realized as a voiceless trill, and the voiced [ʁ] is often realized as an approximant. However, we do not propose any changes at this stage, and acknowledge that the post-velar region of the consonant space is still comparatively poorly understood. Despite the existence of languages (such as in the Caucasus and in parts of the Americas) that treat velar and uvular stops as separate phonemes [15], we note the usefulness of treating velar-uvular as a continuum, varying according to vowel context, or according to language-internal forces [16]. Indeed, Catford's [17] reference to velar and uvular being an "octave" apart can serve as a useful analogy for the gradient nature of this contrast.

On a final short note regarding the velar place of articulation, we wonder if it is entirely accurate to include the velar lateral [ʟ] as a speech sound, given that empirical investigation of Mee has suggested that this sound is highly variable, and could in most cases be characterized simply as the sequence /gl/ [18]. Similar variability has been noted for Mid Waghi and Archi [15, chapter 6]. It is also not clear how the posterior portion of the tongue can maintain a velar closure at the same time as the sides of the tongue allow lateral airflow [cf. 19].

2.3. Retroflex

In terms of place of articulation, the "retroflex" stands out as being particularly odd. This is the one lingual "place" of articulation that is not actually a place – it refers to an idealized tongue configuration (perhaps characterized as sub-apico post-alveolar) that may, or may not, be necessary to produce the particular sounds in this column. We have re-labelled this column as "postalveolar" in order to be more consistent in the naming of the columns, but have kept the same symbols. In our teaching experience, it is difficult to explain the retroflex "place" of articulation and to distinguish it from postalveolar. More importantly, to our ears, it is also almost impossible to hear the difference between a properly retroflex sub-apico postalveolar, and an apical post-alveolar; indeed, a perfectly acceptable "retroflex" sound may be produced simply by retracting the tongue tip into the correct region, without necessarily retroflexing the tongue. Several Indo-Aryan languages such as Hindi, Nepali, and Bengali, which are often described as having retroflex consonants, can be analysed as having retracted alveolars (non-retroflex postalveolars) instead [15, 20, 21]. Similar observations have also been made for Australian Aboriginal languages [29]. For further discussion of variability in retroflex production, the reader is referred to [30].

2.4. Palato-alveolar and alveolo-palatal

There are two "places" of articulation that existed in the 1951 version of the chart that were removed in 1989. A column "palato-alveolar" containing only the fricatives [ʃ ʒ] was removed, as was a column "alveolo-palatal" containing only the fricatives [ɕ ʑ]. The fricatives [ʃ ʒ] were re-labelled as "postalveolar", and are now the only sounds that occur in that column. As we discussed above, the distinction between "postalveolar" and "retroflex" is problematic, since a stop, nasal or lateral sound that is produced at the postalveolar place of articulation, with a retroflexed tongue tip, is auditorily very difficult to distinguish from a sound produced at that same place of articulation, but without a retroflexed tongue tip.

At the same time as [ʃ ʒ] were moved to postalveolar, [ɕ ʑ] were demoted to the Other Symbols category, despite how common these sounds are in many languages of the world, particularly in the languages of East and Southeast Asia. One possibility is to re-introduce the column alveolo-palatal to include [ɕ ʑ], and to also include the stops, nasals and laterals [ȶ ȡ ȵ ȴ], which are used by many scholars of Chinese and other languages. The stop symbols [ȶ ȡ] are also used by authors who consider the palatals [c ɟ] to be more akin to the sequence [kj] rather than the sequence [tj] (i.e. they see the palatal symbols as being more akin to a fronted velar than a properly [alveolo-] palatal sound, the latter being typically associated with extensive affrication of the stops). Whilst the inclusion of an alveolo-palatal place of articulation might be helpful for authors who wish to distinguish [ȶ ȡ] from [c ɟ] along these lines, our main concern is that there is no such auditory distinction between the sounds [ɲ ʎ] and [ȵ ȴ]. We cannot hear a reliable difference between a palatal nasal or lateral, and an alveolo-palatal nasal or lateral. In the case of nasals, this may in part be because the acoustics of nasal consonants are determined by the backmost point of contact for the consonant, rather than the frontmost point of contact, with both a regular palatal and an alveolo-palatal involving a very large degree of tongue-palate contact (although we acknowledge that formant transitions into the vowel do involve the cavity anterior to the constriction) [22].

In the case of laterals, a lack of difference between [ʎ] and [ȴ] may be due to insufficient degrees of freedom in the tongue tip-blade-body complex: the lowering of the jaw and/or narrowing of the tongue required for lateral production may lead to the constraint that the central tip/blade closure cannot be located further back than the pre-palatal region, without leading to retroflexion and the production of /ɭ/ instead of a palatal lateral. In addition, to our knowledge no phonetician has ever proposed an alveolo-palatal glide that is separate from the palatal /j/, yet this could be considered a logical extension to a system that has alveolo-palatal stops, nasals and laterals.

We therefore consider the addition of the alveolo-palatal place of articulation a controversial addition to the chart, and (apart from the fricatives) we think it should at most include the stop manners of articulation. However, even the individual authors of this submission cannot agree on the inclusion of alveolo-palatal stops, and we therefore do not include them in the chart, in large part because their inclusion would involve a re-consideration of the value of the regular palatals [c ɟ]. We return to the issue of how best to describe the contrast between [ʃ ʒ] and [ɕ ʑ] when we consider the overall system of fricatives further below.

3. Fricatives

A particularly difficult aspect of the chart is the fricatives, more precisely the coronal fricatives. The problems can be divided into two categories: one relating to the place-of-articulation distinctions, the other relating to the paradigm of manner differences.

3.1. Problems with place

That place of articulation is a problem for coronal fricatives is shown by the confusion surrounding the terms alveolo-palatal and palato-alveolar (traditionally [ɕ] and [ʃ] respectively), with the latter being re-labelled "postalveolar" in the modern version of the chart. We believe this terminological confusion arises largely because place of articulation is not a sufficient criterion for describing and differentiating these sounds – the extent of grooving (constriction width as well as length) and airflow rate are just as crucial. For instance, in a real-time MRI study of 10 speakers, Yoshinaga et al. [23] found that Japanese palatal [ç] and alveolo-palatal [ɕ] had almost identical places of articulation. Using articulatory modelling, they found that these sounds were differentiated once constriction width and airflow rate were considered.

The question of constriction location, width/length and flow-rate is intimately tied to the question of whether a sound is a sibilant or not. It is well understood that the English fricatives [s z ʃ ʒ] are sibilants, and it has been suggested that an important aspect of their articulation is the central groove directing airflow towards the teeth – the end result being an increased intensity of spectral noise prominence [24]. It is also generally understood that the English fricatives [f v θ ð] are non-sibilants, in the sense that they are not as loud, and in the sense that there is no central groove directing airflow towards the teeth (indeed, this may be impossible for sounds that are labio-dental and dental). Moreover, the sibilant versus non-sibilant distinction is well established in English morpho-phonological rules.

However, when it comes to the other fricatives in the chart, it is not so clear what is sibilant and what is non-sibilant. Yoshinaga et al. [23] initially considered [ɕ] as sibilant and [ç] as non-sibilant based on previous literature. But they subsequently found it difficult to determine an articulatory/acoustic basis for this description, in that the jet of air produced by the constriction reached the edge of the upper incisors in the models for both sounds. This is just one study, and the important point is that there is not the volume of work on non-English fricatives that is needed in order for phoneticians to better understand this class of speech sounds. As expert phoneticians, none of us can confidently say which of the non-English fricatives on the chart is sibilant and which is non-sibilant; we feel that in the absence of a great volume of articulatory and acoustic studies, phonological evidence from a variety of languages is the best evidence we could expect in this regard (but see below for another possible approach to the question of sibilance).

Finally, it might be noted that [ʂ] may or may not involve retroflexion of the tongue tip, as noted above regarding the plosives at the postalveolar place of articulation.

3.2. Problems with manner

The other important problem with the fricatives is that not all fricatives can be derived through articulatory lenition from the corresponding stop place of articulation; and by extension, they cannot themselves be lenited to an approximant at the same place of articulation. If we consider [s], it has a very different tongue shape from [t] (including but not limited to the grooving described above). Indeed, when we speak of a lenited [t] in Australian English, we use the lowering diacritic beneath the stop symbol [t̞]. By extension, one does not speak of /s/ being lenited to /ɹ/. Similar observations could be made for [ʃ ʂ ɕ]. By contrast, the other fricatives operate very well in the stop-fricative-approximant lenition continuum, namely (working with the voiced obstruents for this example) [ɟ ʝ j], [ɡ ɣ ɰ] and [ɢ ʁ]. One would even include dental [d̪ ð ð̞] in this set.

Thus, one could treat the class of sibilants as fricative sounds that cannot be derived by articulatory lenition from a corresponding stop, or strengthening from a corresponding approximant. Indeed, phonological accounts of lenition argue that lenition of stops invariably results in non-sibilant fricatives [25, 26]. Sibilant sounds are produced with a very special tongue configuration that may include significant grooving.

In order to respect the fact that sibilants cannot be derived by articulatory lenition from stops, we mark these sibilant fricatives on the chart with a special double line around the set, in order to offset them. We include the alveolo-palatal and palato-alveolar places of articulation in this new set, as per the pre-1989 chart, but are very conscious that these place labels do not fully describe the articulations of all speakers. For any speech sound, there is a tremendous amount of inter-speaker variability in terms of active articulator used and in terms of precise place of articulation – careful examination of individual data in any articulatory study cannot fail to highlight this. In the case of fricatives, differences between speakers are all the more salient, as this is a class of sounds where the spectral shape of the output noise generated at the constriction is crucial, and the location/shape of this constriction may be highly dependent on individual morphology. However, the acoustic output is highly consistent across speakers despite differences in articulatory input, and as is ultimately the case in all phonetics, it is the acoustic output that is most important.

Finally, it is important to point out that lip rounding plays a role in the production of the palato-alveolars [ʃ ʒ]. This has been remarked upon in many articulatory-to-acoustic modelling studies and even mentioned in textbooks [27, page 159]. This is particularly relevant in the consideration of the difference between these sounds and the alveolo-palatal [ɕ ʑ]: many speakers can produce [ɕ ʑ] as the unrounded version of [ʃ ʒ], despite Catford's [17] suggestion that [ɕ] can be treated as [ʃ] + [j]. The extent to which our knowledge of fricative production (articulatory and acoustic) is imperfect cannot be overstated.

4. Some problems with the manners of articulation

4.1. Approximants

The current chart contains a row for approximants, which includes the semi-vowels [j ɰ], as well as the rhotics [ɹ ɻ] and the labiodental [ʋ]. Curiously, the labio-velar [w], one of the most common consonant sounds in the languages of the world [31], was demoted to "Other Symbols" in the 1989 revision to the chart. Previously it was located in the "Bilabial" column, in the row "Frictionless Continuants and Semi-vowels", sharing a cell with the labio-palatal [ɥ]. [w] also appeared in brackets in the velar column in the same row, and [ɥ] appeared in brackets in the palatal column. Significantly, this row was at the bottom of the chart, closest to the vowel chart, as a signal that the semi-vowels were acoustically, articulatorily and phonologically linked to the vowels. Indeed, the vowel chart pre-1989 was clearly aligned with the palatal and velar columns of the consonant chart, sending a very clear signal of the relationship between the vowels and the palatal and velar approximants.

Conceptually, we would suggest that it would be wise to return to the pre-1989 situation regarding the semi-vowels, especially so given that (with the addition of the velar glide [ɰ]) we now recognize four semi-vowels that can be derived from/related to the four high (corner) vowels – namely [j ɥ ɰ w], derived from [i y ɯ u] respectively. We have therefore adopted this approach in our proposed chart in the Appendix. However, we go further, in that our chart lists a separate row for "Semi-vowels" below the row for approximants. We do this to explicitly show the relationship between the vowels and the semi-vowels – however, we note that the two rows (semi-vowels and approximants) could be combined, and the non-semi-vowels simply offset with a double line, as was done with the sibilant fricatives. This solution would be entirely possible, since (at present) combining the two rows would not result in any overlap of cells. However, if one were to introduce a separate symbol for the lenited bilabial fricative [β̞], as is found in Spanish, then in that case there would be a bilabial derived from vowels (the semivowel [w]) and a bilabial derived from regular lenition (the approximant [β̞]).

Finally in this section, we briefly need to consider the rhotics [ɹ ɻ] under the approximants label. Although we do not propose any changes to these symbols/sounds, we do note that discerning the difference between them is very difficult in practice. The first author simply tells students that [ɻ] sounds darker than [ɹ], presumably reflecting a balance of energy weighted towards the lower part of the spectrum. In principle [ɻ] should be produced further back in the oral cavity than [ɹ], and according to the LAPSyD database, there is only one language that contains a contrast between the two sounds, Wiyot (https://lapsyd.huma-num.fr/lapsyd/index.php?data=view&code=602). It has moreover been suggested that [ɹ] and [ɻ] often involve frication: for example in Beijing Mandarin Chinese the onset/initial postalveolar rhotic varies between approximant and fricative realizations (see discussion in [28]). Indeed, in the previous version of the chart, [ɹ] appeared in both the fricative and the frictionless continuants [i.e. approximants] row. This is another example of where the boundary between fricative and approximant is blurred (in contrast to the boundary between stop and fricative).

4.2. Some notes on other symbols not already mentioned

Here we briefly discuss some individual symbols that strike us as problematic in some way. We do not necessarily suggest that these symbols should be removed – however, we wish to bring these problems to the foreground, so that there is a better appreciation of the inaccuracies and difficulties of the chart.

Firstly, it is curious that there is a special symbol for the voiceless labial-velar fricative [ʍ], which could equally well be represented by [w̥]. We do not suggest that the symbol should be removed, but instead simply note that it is a relic of the history of the IPA chart, which was in its early days heavily influenced by the study of the European languages. What is particularly strange, however, is that the sound is labelled as a fricative, when its counterpart [w] is labelled as an approximant. Does this mean that the noise is supra-laryngeal only? Or is the noise source both glottal and supra-laryngeal? Is it a spread-glottis or breathy-voiced version of the approximant, i.e. [w̤], as found in some North American and south-east Asian languages? Is it simply a sequence of [h] plus [w], as in English? None of this is quite clear.

Another symbol that clearly shows the bias of the early days of the IPA is [ɧ], described as a simultaneous [ʃ] and [x]. Students find this sound very challenging and in our experience most instructors skip over it, saying that it only occurs in Swedish, and that even there it varies greatly dialectally [15]. A sound that occurs in only one language and shows a lot of variation should cause (phonetic) concern, and to our ears the Swedish sound may best be described as a labialized/rounded velar fricative [x̹] or [xʷ] (i.e. with a rounded diacritic beneath the fricative or with a labialization superscript). The symbol [ɧ] is perhaps best removed from the chart since it suggests a sound that does not exist.

Another symbol that is listed under “Other Symbols” is the "alveolar lateral flap" [ɺ]. It is posited as a phoneme in 30 languages in LAPSyD, particularly in languages of South America and Papua New Guinea, where articulatory and acoustic phonetic description is scarce. Notably, however, there are no languages that contrast this sound with the retroflex flap [ɽ] or with the retroflex lateral [ɭ]. Moreover, allthough [ɭ] is not classified as a flap, it is possible to produce this sound as a flap with the *tongue tip* moving forward during closure. To flap the sides of the tongue would involve an unrealistic degree of control (cf. [ʟ] above). It therefore seems that the "alveolar lateral flap" needs to be re-considered.

5. Conclusions

Our proposal for a revised version of the chart is shown in the Appendix. Whilst we hope that it is an improvement on the present chart, we also hope that our discussions in this paper have shown that it is far from perfect. We hope that we have highlighted some of the problems that are inherent to the nature of the IPA chart. It is sometimes assumed that the chart is perfect, and in some way analogous to the chart of chemical symbols or the chart of astronomical objects. However, since the IPA chart is based on human behaviour, this cannot be so. So much about articulation is assumed by a symbol, even when that assumption is idealistic or even at times false. The use of symbols leaves little room for phonetic underspecification, unless the symbols themselves are underspecified (a discussion we have not entered into). We have tried to point out some of the issues with the chart, in the hope that users will not take its theoretical assumptions for granted. As eloquently noted by an anonymous reviewer, "we are just trying to carve up a continuum in one way versus another into categories and there will always be remaining problems and issues".

6. Acknowledgements

We would like to thank Pascal Perrier and Pierre Badin for discussion on issues of tongue control, and our anonymous reviewers for crucial feedback.

7. References

[1] International Phonetic Association, "Report on the 1989 Kiel Convention". Journal of the International Phonetic Association, *19,* 67-80, 1989.

[2] Gordon, M., Syllable weight: phonetics, phonology, typology, New York: Routledge, 2006.

[3] Garellek, M., Chai, Y., Huang, Y., & Van Doren, M., "Voicing of glottal consonants and non-modal vowels", Journal of the International Phonetic Association, 53(2), 305-332, 2023.

[4] Ladefoged, P., "Some proposals concerning glottal consonants", Journal of the International Phonetic Association, 20(2), 24-25, 1990.

[5] Esling, J. H., Moisik, S. R., Benner, A., & Crevier-Buchman, L., Voice quality: The laryngeal articulator model (Vol. 162). Cambridge University Press, 2019.

[6] Udomkesmalee, N., "Historical charts of the International Phonetic Alphabet", 2018. https://www.internationalphoneticassociation.org/IPAcharts/IPA_hist/IPA_hist_2018.html

[7] Clements, G. N., "The geometry of phonological features", Phonology, 2, 225-252, 1985.

[8] McCarthy, J., "Feature geometry and dependency: a review", Phonetica 45, 84-108, 1988.

[9] Robins, R. H., "Vowel nasality in Sundanese: a phonological and grammatical study", in J. R. Firth (ed.), Studies in Linguistics, 87-103. Oxford: Basil Blackwell, 1957.

[10] Cohn, A., Phonetic and phonological rules of nasalization, PhD dissertation, UCLA. (UCLA Working Papers in Phonetics 76), 1990.

[11] Bennett, R., "Mayan phonology", Language and Linguistics Compass 10(10), 469-514, 2016.

[12] Rogers, C., "Vowel harmony in Meso-American languages", 691-699. In N. Ritter & H. van der Hulst (eds.), The Oxford handbook of vowel harmony, 691-699. New York: Oxford, 2014.

[13] Stemberger, J., "Glottal transparency", Phonology 10, 107-138, 1993.

[14] Catford, J. C., "Glottal consonants… another view", Journal of the International Phonetic Association, 20(2), 25-26, 1990.

[15] Ladefoged, P. and Maddieson, I. The sounds of the world’s languages. Oxford: Blackwell, 1996.

[16] Butcher, A. & Tabain, M. "On the back of the tongue: dorsal sounds in Australian languages", Phonetica 61 22-52, 2004.

[17] Catford, J. C. A practical introduction to phonetics (2nd edition). New York: Oxford University Press, 2001.

[18] Staroverov, P. & Tebay, S., "Velar lateral allophony in Mee (Ekari)", Journal of the International Phonetic Association, 52, 278–308, 2022.

[19] Perrier, P., Payan, Y., Zandipour, M. & Perkell, J. "Influences of tongue biomechanics on speech movements during the production of velar stop consonants: a modelling study", Journal of the Acoustical Society of America, 114, 1582–1599, 2003.

[20] Khatiwada, R., "Nepali", Journal of the International Phonetic Association, 39(3), 373-380, 2009.

[21] Khan, S. D., "Bengali (Bangladeshi Standard)", Journal of the International Phonetic Association, 221-225, 2010.

[22] Fant, G., Acoustic Theory of Speech Production (2nd edition). The Hague: Mouton, 1970.

[23] Yoshinaga, T., Maekawa, K. & Iida, A., "Aeroacoustic differences between the Japanese fricatives [ɕ] and [ç]", Journal of the Acoustical Society of America, 149, 2426–2436, 2021.

[24] Shadle, C., The acoustics of fricative consonants, PhD Thesis: MIT, 1985.

[25] Kirchner, R., An effort-based approach to consonant lenition, Ph.D. dissertation, University of California, Los Angeles, 1998.

[26] Kirchner, R., "Consonant lenition", in B. Hayes, D. Steriade, and R. Kirchner (eds.), Phonetically based phonology, 313-345. New York: Cambridge University Press, 2004.

[27] Johnson, K., Acoustic and Auditory Phonetics (3rd edition), Maldon MA, Oxford, West Sussex: Wiley Blackwell, 2012.

[28] Xing, K., Phonetic and phonological perspectives on rhoticity in Mandarin, Ph.D. thesis, University of Manchester, 2021.

[29] Butcher, A., The sounds of Australian languages, Unpublished manuscript.

[30] Hamann, S., The Phonetics and Phonology of Retroflexes, Ph.D. dissertation, Utrecht: LOT Press, 2003.

[31] Gordon, M. Phonological Typology, Oxford: OUP, 2016.

Appendix: *Proposed revision to IPA Consonant Chart.*

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Palato-alveolar	Alveolo-palatal	Palatal	Velar	Uvular	Epilaryngeal Pharyngeal	Laryngeal
Plosive	p b			t d	ʈ ɖ			c ɟ	k g	q ɢ	ʡ	ʔ
Nasal	m	ɱ		n	ɳ			ɲ	ŋ	ɴ		
Trill	ʙ			r						ʀ	ʜ ʢ	
Tap/flap		ⱱ		ɾ	ɽ							
Fricative	ɸ β	f v	θ ð	s z	ʂ ʐ	ʃ ʒ	ɕ ʑ	ç ʝ	x ɣ	χ ʁ	ħ	
Lateral fricative				ɬ ɮ								
Lateral Approximant				l	ɭ			ʎ	ʟ			
Approximant		ʋ		ɹ	ɻ						ʕ	h ɦ
Semi-vowels	w | ɥ							j (ɥ)	ɰ (w)			

Table of Contents for this manuscript

Lateral Articulation Across Vowel Contexts: Insights from a Magnetic Resonance Imaging Case Study

Tünde Szalay[1], Michael Proctor[1], Amelia Gully[2], Tharinda Piyadasa[3], Craig Jin[3], David Waddington[4], Naeim Sanaei[5], Yaoyao Yue[3], Sheryl Foster[4,5], Kirrie Ballard[6]

[1]Dept. of Linguistics, Macquarie University,
[2]Dept. of Language and Linguistic Science, University of York,
[3]School of Electrical and Computer Engineering, University of Sydney,
[4]Image X Institute, University of Sydney, [5]Radiology Department, Westmead Hospital,
[6]Discipline of Speech Pathology, University of Sydney

`tunde.szalay@mq.edu.au`

Abstract

The goals of lateral production are complex and imperfectly understood, partly because of the limitations of existing data. Structural Magnetic Resonance Imaging provides rich information about details of lateral production not available using other methods. /l/-articulation by a British English speaker was examined in three vowel contexts using real-time and volumetric Magnetic Resonance Imaging. Onset of lateralisation was characterized acoustically by decreased intensity and development of anti-formants, independent of the degree of tongue dorsum retraction and lingual elongation in different coarticulatory contexts. These patterns suggest that, for this speaker, active lateral channel formation is a primary goal of clear-/l/ production.

Index Terms: liquids, approximants, articulatory-acoustic relationships, coarticulation, goals of production, rtMRI

1. Introduction

English /l/ is a multigestural segment prototypically produced with a central alveolar closure, dorsal retraction and lowering, and lateral channel formation [1, 2]. Lateral approximants are characterised by complex intergestural and articulatory-acoustic relationships [3, 4, 5]. Lateral channels may form passively when the tongue is elongated through simultaneous tongue tip fronting to achieve alveolar closure and dorsal retraction, as is typically observed in dark [ɫ] [6]. Lateral channels also form in clear [l] articulated with less lingual elongation, where stable timing relations have been observed between the sides and back of the tongue, suggesting that there may be active control of lateralisation [7, 8]. Many details of lateral production are still not well understood, in part due to the limitations of methods used to study the configuration of the vocal tract.

Acoustic data offer important insights into lateral production, as /l/ typically shows three distinct formants below 5 kHz [3, 5, 9]. The low F1 (∼ 250–500 Hz) is associated with a Helmholtz resonance between the relatively large back cavity volume and the oral constriction space [3, 5]. F1 increases when the oral constriction is reduced, contributing to the higher F1 in dark [ɫ] produced with a weakened coronal contact [3, 10]. F2 (∼ 1.2–1.5 kHz) is associated with the back cavity, such that retracting or raising the tongue dorsum increases back cavity length and lowers F2 in dark [ɫ] [3]. Lateralized airflow can give rise to spectral zeros whose properties depend on the length and asymmetry of the channels; anti-resonances >3 kHz can result when a pocket of air above the tongue forms a side branch to the primary lateralized airway [9, 11]. Anti-resonances raise the 3rd formant, and a high F3 well separated from F2 is one of the defining acoustic features of lateral approximants [5, 11, 12].

Lateral production has been studied using sustained /l/ [3, 4] and in specific vowel contexts [5, 10], yet dorsal posture – a key gesture affecting tongue elongation and F2 – varies with vowel-context, assuming a similar articulatory target to that of adjacent vowels [13]. In American English, onset [l] is coarticulated more strongly with the vowel than coda [ɫ] [13], predicting considerable coarticulatory F2 variation; however, Catalan /lˠ/ is articulated with a lower dorsum between low vowels, and is produced with a relatively stable F2 across vowel contexts [14].

To further examine these relationships, we analyzed time-aligned articulatory and acoustic data in a single speaker study of Standard Southern British English (SSBE) /l/ produced in three vowel contexts, using real-time (rtMRI) and volumetric magnetic resonance imaging (MRI). Our aims are to (1) identify articulatory and acoustic /l/ targets; (2) describe the coarticulatory influences of vowel context on /l/; and (3) link the articulatory changes caused by vowel context to acoustic changes.

2. Methods

Data were collected during the pilot phase of a larger project examining development of speech motor control in adolescents. An adult female L1 speaker of SSBE (Author 3) produced intervocalic laterals in a series of speech tasks recorded out of and inside an MRI scanner. Laterals were elicited between three corner vowels: high front /iː/, low /ɑː/, and high back /uː/. Each token was recorded once in a quiet room with a Glottal Enterprises EG2-PCX2 digital speech recorder to familiarize the participant with the experimental materials. The same utterances were later recorded three times during a rtMRI scan, and additionally as sustained lateral productions during a volumetric MRI scan. A total of 3 (vowel contexts) × (1 pre-scan + 3 rtMRI) + 1 (volumetric MRI) = 13 laterals were included in the analysis.

2.1. Data acquisition

MRI data were acquired at Westmead Hospital (Sydney, New South Wales), on a Siemens Magnetom Prisma 3T scanner with a 64-channel head/neck receiver array coil. The speaker's upper airway was imaged while lying supine. Data were acquired from an 8 mm slice aligned with the mid-sagittal plane, over a 280 × 280 mm field of view, using a 2D RF-spoiled, radially-encoded FLASH sequence [15]. Audio was recorded concurrently in-scanner at 16 kHz using an Opto-acoustics FOMRI-III ceramic noise-canceling microphone designed for MRI environments [16]. rtMRI data were reconstructed in Matlab into midsagittal videos with a pixel resolution of 0.83 mm^2, encoded as 72 frames per second MP4 files. Audio and video were time-aligned during postprocessing and video reconstruction based on visual inspection of the audio signal and the video frames.

3D configuration of the vocal tract during sustained (7.6 s) lateral production was captured using volumetric imaging of the upper airway. Data were acquired using a T1-weighted fast 3D gradient-echo sequence, with a spatial resolution of 160 × 160 × 32 px over a 256 × 256 × 64 mm field of view centred on the pharynx: a voxel resolution of 1.6 × 1.6 × 2.0 mm.

2.2. Phonetic data analysis

rtMRI videos and time-aligned in-scanner audio recordings were analyzed using a Matlab-based custom graphical interface. Image frames were identified corresponding to articulatory target postures for pre- and post-lateral vowels, and lateral coronal closure, target, and coronal release (Figs. 1, 3, 6).

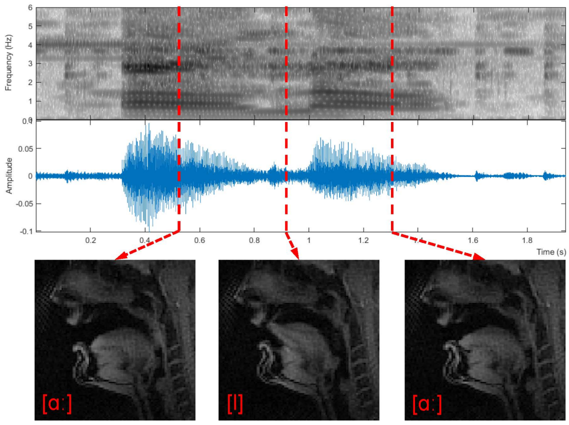
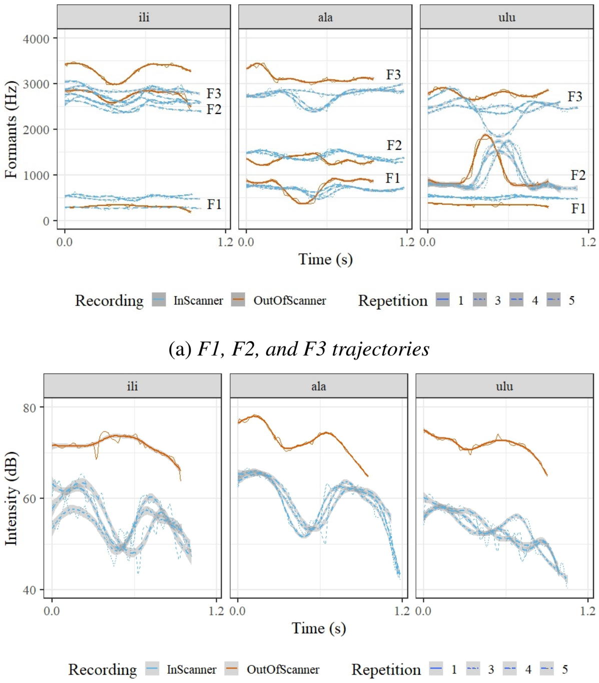


Figure 1: *Intervocalic lateral production, low vowel context.* Spectrogram and waveform of noise-cancelled in-scanner recording of /ɑlɑ/, time-aligned with rtMRI frames captured at vowel and lateral lingual target postures.

Vowel targets were located at the centre frame of the stable articulatory position associated with each segment (Fig. 1, bottom L, R). Lateral coronal closure was located at the first frame after any observable gap between the tongue tip (TT) and alveolar ridge (Fig. 6, bottom centre L). The lateral target was located at the centre frame of the interval over which contact was maintained between the TT and alveolar ridge (Fig. 1, bottom centre). Lateral coronal release was located at the first frame when a gap between the TT and alveolar ridge was first observed after closure (Fig. 6, bottom centre R). Coronal closure was achieved in every token, showing that none of the laterals were vocalised. Lingual target postures were identified in all tokens despite the motion blur in some frames, as slow and hyperarticulated speech yielded visible sustained lingual targets.

Audio recordings were force-aligned using MAUS to locate segment boundaries which were then hand-corrected [17, 18, 19]. Formant trajectories were estimated automatically and corrected manually in Praat [20]. Formant frequencies were estimated every 10 ms over a 40 ms Gaussian analysis window with 75% overlap, 50 dB dynamic range, and a pre-emphasis filter increasing spectral slope above 100 Hz by 6 dB/octave. Five formants were tracked up to a 5.5 kHz ceiling for tokens with higher F2 values, and up to 5 kHz for lower F2 values, then corrected manually [21]. At each timepoint where formants were estimated, intensity values were estimated with a pitch floor of 100 Hz. Characteristic formant trajectories and intensity contours for laterals in each vowel context were generated by fitting a Generalised Additive Model (GAM) to the set of time series for each experimental item (Fig. 2).



(a) *F1, F2, and F3 trajectories*

(b) *Intensity contours*

Figure 2: *Formant trajectories (a) and intensity contours (b).* Out-of-scanner (red) and 3 in-scanner (blue) repetitions of (L-to-R): [iːliː], [ɑːlɑː], [uːluː]. Individual repetition timeseries (thin lines) fitted with GAMs (thick lines + grey s.d.).

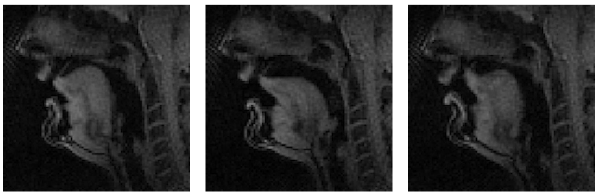
2.3. Acoustic data validation

Formant and intensity values estimated in in-scanner recordings were validated against acoustic measures estimated in out-of-scanner recordings. Formants generally tracked poorly in in-scanner recordings due to scanner noise and signal processing involved in noise-reduction. F1 and F2 estimates from in- and out-of-scanner recordings aligned most closely, but F3 estimates were consistently lower for in-scanner recordings, compared to out-of-scanner equivalents, and showed larger discrepancies between repetitions (Fig. 2a). Manual correction of 3rd formant trajectories in these data was determined to be too unreliable, so F3 was not analysed for in-scanner recordings.

Utterance durations were consistently longer for in-scanner recordings, primarily due to lengthening of pre-lateral vowels, but the same general patterns can be observed as the speaker hyperarticulated during the rtMRI scan (Fig. 2). Vowel and lateral intensity was consistently higher in out-of-scanner recordings compared to in-scanner recordings (Fig. 2b). Intervocalic laterals showed an intensity dip relative to the initial vowel in all tokens other than out-of-scanner /ili/ (Fig. 2b), which is attributed to speech variation, rather than vowel or recording environment effects. Laterals, however, only showed lower intensity relative to the final vowel in [ili] and /ɑlɑ/, but not in /ulu/, due to /i/ and /ɑ/ having higher intensity than /u/ (Fig. 2b).

3. Results and Discussion

Complete midsagittal occlusion in the dental-alveolar region was observed in all lateral tokens (Fig. 3); no vocalized /l/ was produced in these data. Achieving tongue tip contact in intervocalic /l/ is consistent with the lack of undershoot in this position in American- and British English [1, 22]. Extended /l/ duration could also contribute to achieving alveolar closure by providing enough time for the tongue tip to reach its target [23, 24].



(a) [iːliː] (Rep. 4) (b) [ɑːlɑː] (Rep. 4) (c) [uːluː] (Rep. 4)

Figure 3: *Midsagittal /l/ articulation at TT target in three vowel contexts.* L-to-R: [iːliː], [ɑːlɑː], [uːluː].

Volumetric image data show the 3D vocal tract configurations used to produce laterals in each vowel context, including key details of /l/ articulation beyond the midsagittal plane (Fig. 4). The central occlusion formed by the TT against the alveolar ridge can be seen anterior to the mid-oral cavity. On either side of the occlusion, narrow lateral channels connect the mid-oral airway to the anterior part of the vocal tract formed by the sub-lingual cavity. The precise geometry of the lateral channels cannot be determined from this volume because these regions will also include some dentition (teeth do not image in MRI); yet, the data reveal two largely symmetrical lateral channels and a complete central alveolar occlusion.

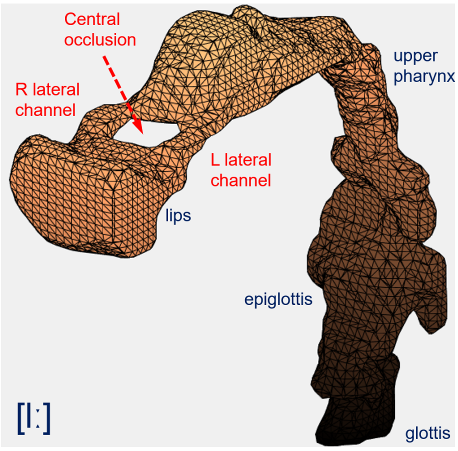


Figure 4: *Three dimensional vocal tract configuration during sustained [lː].* Tract volume viewed from L. superior anterior perspective. Anterior part of volume extends beyond lips.

3.1. Acoustic characterization of intervocalic laterals

Laterals were elicited in intervocalic environments, where they were acoustically delineated by a drop in intensity (Fig. 2b), formant transitions specific to the vowel context (Fig. 7), and appearance of anti-formants. Intensity drop cued lateral onset and offset in the [iː] and [ɑː] contexts and lateral onset in the [uː] context (Fig. 2b). A spectrogram generated from the pre-scan recording of [ɑːlɑː] (Fig. 5) reveals a prominent antiresonance centred at 3.7 kHz throughout the lateral interval (690 to 920 ms). In Fant's model, both reduced intensity and antiresonance would arise from a side branch of length ∼23 mm [9, 25, 26]. Although the precise length of the supralingual air pocket cannot be determined because of uncertainties associated with dentition, the frequencies of the main antiformants observed in these spectra are broadly consistent with Fant's acoustic model applied to the vocal tract configurations revealed by the imaging data. Anti-resonances in a similar region were observed in American English /l/, while intensity drop characterises Turkish and Brazilian Portuguese laterals [4, 27].

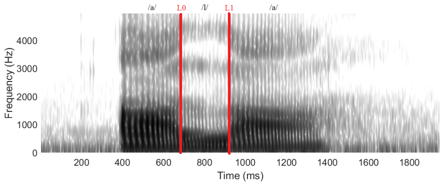


Figure 5: *Spectrogram of [ɑːlɑː] (out-of-scanner): 6 ms Kaiser windows, 2 ms overlap, 1024 pt FFT.* L0: beginning of lateral; L1: end of lateral. Primary anti-formant centred at 3.7 kHz.

3.2. Vocalic influences on lateral production

Imaging data reveal large coarticulatory influences of vowel context on lateral production. Coronal place of articulation varies in anteriority with vowel frontness: dental-alveolar for [iːliː] (Fig. 3a) and alveolar for [ɑːlɑː] (Fig. 3b), both produced with an apical TT gesture. In [uːluː], the midsagittal constriction occurs at a more retracted post-alveolar target through sub-laminal TT closure and a more retroflexed coronal gesture (Fig. 3c). The dorsum is raised and fronted in the high-front vowel context (Fig. 3a), lowered in the low vowel context (Fig. 3b), and high and back in the back vowel context (Fig. 3c).

F1 in intervocalic laterals ranged from 350 Hz in high vowel contexts to 750 Hz between low vowels (Fig. 7). F1 trajectories are relatively stable throughout [iːliː] and [uːluː], consistent with the stability in tongue height observed in the correspond-

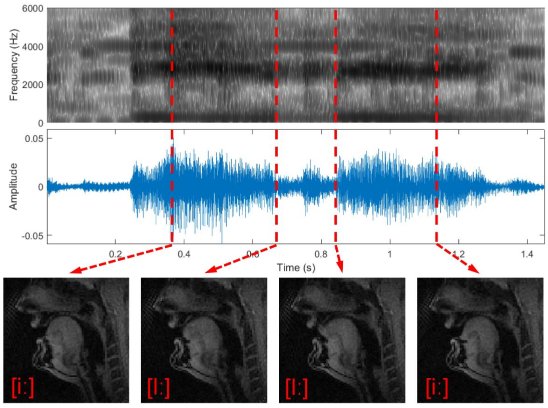


Figure 6: *Lateral dynamics in the front vowel context.* Spectrogram and waveform of noise-cancelled in-scanner recording of /ili/, time-aligned with MRI frames captured at (L to R): pre-/l/ vowel target, /l/ onset, /l/ target, post-/l/ vowel target.

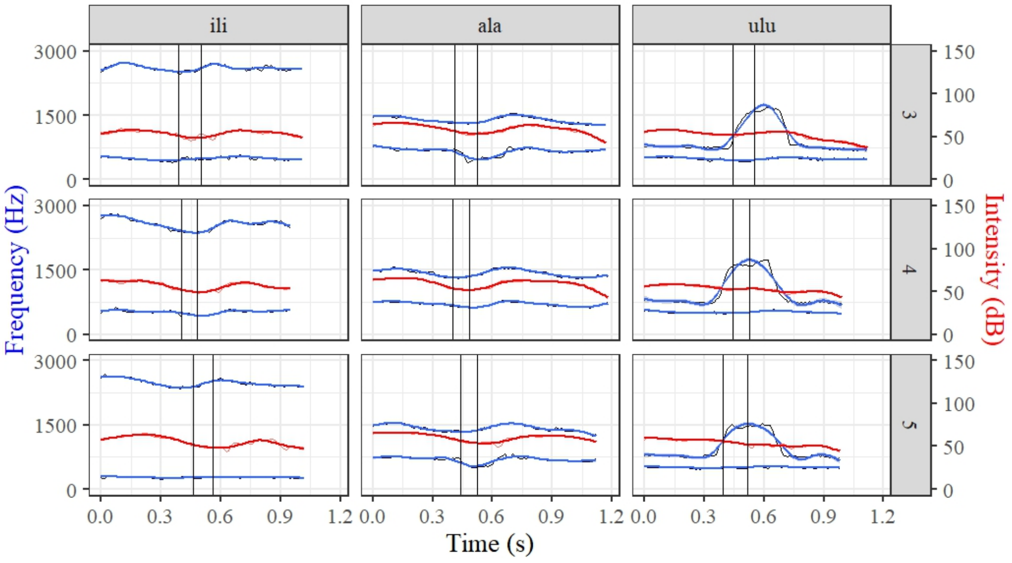


Figure 7: *Formant trajectories (left y-axis, blue) and intensity contours (right y-axis, red), aligned with onset and offset of coronal closure (vertical black lines).* Three in-scanner repetitions of (L to R): /ili/, /ala/, /ulu/. Top to bottom: Repetitions 1 to 3. Formant trajectories and intensity contours (thin lines) smoothed using GAMs (thick lines, std. dev. in grey).

ing image sequences of these utterances (Fig. 6). F1 was higher throughout [ɑːlɑː] utterances, lowering in the transition from the initial vowel to the lateral, then rising to a peak of ∼750 Hz at lateral TT release (Fig. 7) – higher than F1 values previously reported for sustained laterals (350–450 Hz) [3]. The raised F1 in [ɑːlɑː] is consistent with the pervasive tongue body lowering observed in the corresponding midsagittal image sequences of laterals in low vowel contexts (Fig. 1).

A wide range of F2 frequencies are seen in these data. Overall, F2 trajectories show the expected correlations with tongue frontness. In non-front vowel contexts, lateral F2 ranged between 1450–1550 Hz. F2 was most stable throughout [ɑːlɑː] utterances, lowering slightly before TT closure and peaking after TT release, following a similar trajectory to F1 (Fig. 7). [iːliː] was also characterized by F2 lowering in the pre-lateral vowel, however, F2 did not reach the same target in the lateral, remaining much higher (>2 kHz) than in the other vowel contexts (Fig. 7). Between back vowels, F2 rose sharply to peak at ∼1500 Hz at the point of TT closure, then relowering after TT release to the F2∼750 Hz of context [uː] (Fig. 7). Variation in lateral formants is consistent with lateral formants not being sufficient for distinguishing laterals from other segments in SSBE, similarly to other languages (e.g., Turkish, Brazilian Portuguese, Central Australian languages) [28, 27].

These formant trajectories are consistent with the coarticulatory patterns observed in imaging data. Lateral F2 is affected more strongly by adjacent /iː/, compared to /ɑː/ and /uː/. Midsagittal images reveal that the tongue body is more advanced throughout [iːliː] compared to the other vowel contexts, and does not retract as much at the lateral target (Figs. 1, 6). Stronger coarticulatory influences of front vowels, and palatals more generally, have also been demonstrated in Catalan and other languages [29]. As a result of this coarticulation, this speaker's laterals are not consistently produced with an elongated tongue, so it appears unlikely that lateral channels are formed passively in front vowel contexts [6]. These data – although limited – lend more support for models proposing active lateral channel formation in /l/ production [7, 8]. The same patterns of production might also arise if tongue blade width were an active parameter of control [30], allowing side channels to form around a narrowed TT central constriction.

4. Conclusions and future research

The dataset demonstrates the value of multi-modal data in the phonetic characterization of complex segments. Understanding dynamic patterns of articulation beyond the midsagittal plane and their acoustic consequences is particularly important for lateral approximants. This speaker consistently produced hyperarticulated intervocalic laterals with central TT closure and formation of lateral channels, characterized acoustically by antiformant(s) and reduced intensity relative to vowels. Tongue body anteriority during lateral production and formant trajectories – especially F2 – were strongly influenced by vowel context, and may provide less consistent cues to lateralization. Inconsistent tongue body retraction across vowel contexts suggests that active lateral channel formation, rather than lingual elongation, is a primary goal of /l/ production for this speaker.

The data are limited in scope, as only a small number of lateral exemplars from a single speaker of SSBE have been analyzed, and the speech is hyperarticulated due to the nature of the task and the unusual environment in which it was produced. More detailed analysis of the geometry and dynamics of lateral channel formation in different phonological environments is required to better understand how lateralization is achieved, and how /l/ can be characterized in articulatory and acoustic domains. Dynamic imaging in the coronal plane and modelling of dentition in MRI data will help inform these issues. Robust tracking of F3 in in-scanner recordings will be important to better characterize the acoustic dynamics of lateral production.

5. Acknowledgements

This work was supported by Australian Research Council Discovery Grant DP220102933.

6. References

[1] S. B. Giles and K. L. Moll, "Cinefluorographic study of selected allophones of english /l/," *Phonetica*, vol. 31, no. 3-4, pp. 206–227, 1975.

[2] M. Stone, A. Faber, L. J. Raphael, and T. H. Shawker, "Cross-sectional tongue shape and linguopalatal contact patterns in [s],[ʃ], and [l]," *Journal of Phonetics*, vol. 20, no. 2, pp. 253–270, 1992.

[3] S. S. Narayanan, A. A. Alwan, and K. Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals," *The Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1064–1077, 1997.

[4] X. Zhou, C. Y. Espy-Wilson, M. Tiede, and S. Boyce, "An MRI-based articulatory and acoustic study of lateral sound in American English," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4182–4185.

[5] S. Charles and S. M. Lulich, "Articulatory-acoustic relations in the production of alveolar and palatal lateral sounds in Brazilian Portuguese," *The Journal of the Acoustical Society of America*, vol. 145, no. 6, pp. 3269–3288, 2019.

[6] C. P. Browman and L. M. Goldstein, "Gestural syllable position effects in American English," in *Producing speech: contemporary issues (for Katherine Safford Harris)*, F. Bell-Berti and L. J. Raphael, Eds. New York: AIP Press, 1995, pp. 19–34.

[7] M. Proctor, "Towards a gestural characterization of liquids: Evidence from Spanish and Russian," *Laboratory Phonology*, vol. 2, no. 2, pp. 451–485, 2011.

[8] J. Ying, J. A. Shaw, C. Carignan, M. Proctor, D. Derrick, and C. T. Best, "Evidence for active control of tongue lateralization in Australian English /l/," *Journal of Phonetics*, vol. 86, p. 101039, 2021.

[9] G. Fant, *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. s'Gravenhage: Mouton, 1960.

[10] R. Sproat and O. Fujimura, "Allophonic variation in English /l/ and its implications for phonetic implementation," *Journal of Phonetics*, vol. 21, no. 3, p. 291–311, 1993.

[11] K. N. Stevens, *Acoustic Phonetics*. Cambridge, Mass; London: MIT Press, 2000.

[12] C. Y. Espy-Wilson, "Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in American English," *The Journal of the Acoustical Society of America*, vol. 92, no. 2, pp. 736–757, 1992.

[13] M. Proctor, R. Walker, C. Smith, T. Szalay, S. Narayanan, and L. Goldstein, "Articulatory characterization of English liquid-final rimes," *Journal of Phonetics*, vol. 77, p. 100921, 2019.

[14] D. Recasens, M. D. Pallarès, and J. Fontdevila, "Co-articulatory variability and articulatory–acoustic correlations for consonants," *International Journal of Language & Communication Disorders*, vol. 30, no. 2, pp. 203–213, 1995.

[15] A. J. Kennerley, D. A. Mitchell, A. Sebald, and I. Watson, "Real-time magnetic resonance imaging: mechanics of oral and facial function," *British Journal of Oral and Maxillofacial Surgery*, vol. 60, no. 5, pp. 596–603, 2022.

[16] Optoacoustics Ltd., "FOMRI-II version 2.2," 2007.

[17] F. Schiel, "Automatic Phonetic Transcription of Non-Prompted Speech," in *Proc. 14th Intl. Congress of Phonetic Sciences*, J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey, Eds., San Francisco, CA, USA, 1999, p. 607–610.

[18] ——, "A statistical model for predicting pronunciation." in *ICPhS*, 2015.

[19] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, p. 326–347, 2017.

[20] P. Boersma and D. Weenink, "Praat 6.4.05." 2024. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[21] T. Szalay, T. Benders, F. Cox, and M. Proctor, "Vowel merger in Australian English lateral-final rimes: /æo-æ/," in *Proc. 18th Australasian Intl. Conf. on Speech Science and Technology*, R. Billington, Ed., 2022.

[22] J. M. Scobbie and M. Pouplier, "The role of syllable structure in external sandhi: An EPG study of vocalisation and retraction in word-final english/l," *Journal of Phonetics*, vol. 38, no. 2, pp. 240–259, 2010.

[23] E. Lawson and J. Stuart-Smith, "The effects of syllable and sentential position on the timing of lingual gestures in /l/ and /r/," in *International Congress of Phonetic Sciences (ICPhS 2019)*, Melbourne, Australia, 2019, pp. 547–551.

[24] P. Strycharczuk, J. M. Scobbie, A. Przewozny, C. Viollain, and S. Navarro, "Gestural delay and gestural reduction. articulatory variation in/l/-vocalisation in southern british english," *The corpus phonology of English: Multifocal analyses of variation*, pp. 9–29, 2020.

[25] Z. Zhang and C. Y. Espy-Wilson, "A vocal-tract model of american english/l," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1274–1280, 2004.

[26] K. Johnson, *Acoustic and Auditory Phonetics*, ser. Acoustic and Auditory Phonetics. Malden, MA: Wiley-Blackwell, 2011.

[27] S. Charles, "Articulation and acoustics of lateral speech sounds," Ph.D. dissertation, Indiana University, 2022.

[28] M. Tabain, A. Butcher, G. Breen, and R. Beare, "An acoustic study of multiple lateral consonants in three Central Australian languages," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 361–372, 2016.

[29] D. Recasens and C. Rodríguez, "A study on coarticulatory resistance and aggressiveness for front lingual consonants and vowels using ultrasound," *Journal of Phonetics*, vol. 59, pp. 58–75, 2016.

[30] C. P. Browman and L. M. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, no. 2, pp. 201–251, 1989.

Māori /r/ Acoustics: Preliminary Analysis of the First and Second Formant

Isabella Shields, Peter J. Keegan, Catherine I. Watson

University of Auckland

isabella.shields@auckland.ac.nz, p.keegan@auckland.ac.nz, c.watson@auckland.ac.nz

Abstract

This paper describes a preliminary investigation into first and second formant (F1 and F2) behaviours of Māori /r/. A total of 607 observations of this sound are drawn from recordings of 11 fluent Māori speakers. The influence of segmental environment, word form, and phrasal context are explored using mixed-effect models. We find segmental environment impacts /r/ F1 and F2, with both preceding and following vowels modifying these formants. Word form impacts F1 and F2 only in certain segmental environments, and phrasal context has no effect. These findings provide a starting point for further investigation into Māori /r/ and coarticulation.

Index Terms: Māori phonetics, rhotics, acoustics

1. Introduction

Māori is the indigenous language of *Aotearoa* (a common Māori name for New Zealand). Aside from English, Māori is the second-most spoken language in Aotearoa: in 2018, 185,955 people reported they could hold a conversation about everyday things [1]. Colonisation of Aotearoa in the 1800s instigated a period of Māori language decline that was exacerbated by discouragement of use of the language. This resulted in a break in inter-generational language transmission and a decline in the number of speakers. The latter half of the 20th century saw a significant push to revitalise Māori, led predominantly by Māori communities who have, for example, spearheaded the development of Māori-medium and immersion schooling. Revitalisation efforts have also been supported by academic research.

The present research contributes to the overarching goal of improving our knowledge of Māori phonetics. Presently, we do not understand the extent of variation of Māori /r/. It is unclear, for example, what variants are considered 'acceptable' or as an error. This paper reports on a preliminary investigation into the formant behaviours of /r/, and considers the influence of word form, phrasal environment, and surrounding vowels.

2. Background

The phoneme inventory of Māori has 10 consonants: /p t k m n ŋ f h w r/ and five monophthongs, /i e a o u/, which have phonemically distinctive short and long qualities. In written Māori, these long vowels are usually denoted using a macron. Māori is described as a mora-timed language; mora consist of a short monophthong and an optional consonant (μ = (C)V).

Biggs' stress rules describe primary stress placement in phrases and monomorphemic words [2]. While there has been limited research into acoustic evidence of these stress rules, they have been the subject of some investigation [3], [4]. According to these rules, word stress is placed according to two factors; a hierarchy of syllables (summarised by Bauer [5]), and the number of morae in a word. These rules are not expanded upon here, but did influence the choice and form of the target words from which data is extracted in the present study. According to Biggs' rules, phrase stress is assigned based on the position of the phrase in the sentence, and may only be placed on content words. If the phrase is sentence-final, phrase stress falls on stressed syllable of the final content word. If it is not sentence-final, phrase stress falls on the penultimate mora of the phrase.

Māori /r/ is most often described as a flap [5]–[7]. Bauer provides possibly the most detailed description, calling the sound a *voiced lamino-alveolar tap* or *apico-alveolar tap* [5]. In [8], Hohepa describes trill productions of /r/, and Wilson and Harlow refer to lateral realisations of the phoneme [7], [9]. Outside of select contexts, such as when /r/ appears repeatedly with intervening unstressed vowels (e.g. in *kōrero*), Harlow notes that the presence of an approximant variant is a case of interference of New Zealand English with Māori [7].

The Māori rhotic has been the subject of phonetic analysis, most of which has focused on the MAONZE corpus [10], [11]. This corpus contains recordings of 62 bilingual Māori speakers whose birth dates span over 100 years. Maclagan and King undertook an auditory analysis of /r/ as produced by one of the historical speakers (born in 1885 and recorded in 1947) in the MAONZE corpus [10]–[12]. Their work investigated the speaker's Māori and English recordings, finding that he consistently used taps when speaking in Māori, with some approximant productions when he used Māori words in predominantly English speech. Maclagan et al. also considered the pronunciation of the word *māori* in the wider MAONZE corpus [13]. They found some indication of changes over time, with the proportion of taps used in Māori speech decreasing and the proportion used in English speech increasing.

Analysis of other corpora has also been undertaken. In an investigation of three present-day speakers from a new corpus, Shields et al. found the majority of /r/ tokens were indeed taps, and that its phonetic realisation varied [14]. They observed most tokens had at least some formant energy present throughout its production. Frication was present in a number of tokens (more often near /i/), and a smaller proportion also had release bursts. A recent study, analysing MAONZE recordings and more contemporary sources, considered the spectrographic realisation of /r/, the interaction between /r/ duration and stress, and the behaviour of the fourth formant [15].

This paper follows on from the above studies of Māori /r/ and hones in on its first and second formant (F1 and F2) behaviours. This is an area that has not yet been investigated in depth or in a wider speaker group. The present work focuses on realisations of the sound that have formant energy throughout their production, a variant identified in [14] as the most common. Formant values are considered at a single point in time in various word stress, phrase stress, and preceding/following segmental environments. We expect that these variables will have varying degrees of interaction with F1 and F2.

3. Methodology

The present study draws on a corpus of read Māori speech developed to investigate the /r/ phoneme. A preliminary analysis of this corpus was presented in [14] when it was in the early stages of development. The experimental procedure following in this study was approved by the University of Auckland Human Participants Ethics Committee (UAHPEC23198).

3.1. Kaikōrero (speakers) and materials

We present data drawn from the 11 female speakers recorded, all of whom are fluent speakers of Māori. The mean age of speakers was 35.1 years (s.d. = 11.8; range = 17 to 60 years). The oldest female speaker recorded did not explicitly state their age, indicating they were '50+' years. Their age is not included in the reported mean, although they are no older than 60 years. The majority of the speakers grew up with exposure to Māori: most of the speakers (7/11) had at least one Māori-speaking parent, and all had at least one grandparent who spoke or understood Māori. Around half of the speakers (6/11) use Māori at home as an adult, and all speakers use Māori (to varying degrees) at work or university.

In the corpus, two different frame sentences, presented below, place the /r/ token in different phrasal stress conditions (according to rules discussed in Section 2). In frame sentence 2, phrase stress is conflated with word stress of the target word. Phrase stress falls outside the target word in frame sentence 1.

(1) *I/Ka* {target word} *a* {Name}.

(2) *I kite a* {Name} *i te/ngā* {target word} *nā*.

In the present study we analyse tokens taken from three of the five word forms included in the corpus. These word forms place /r/ in different lexical stress environments and segmental environments. The word forms are as follows; WF1 /ˈCV.rV/ (such as *para*), WF2 /ˈCVː.CV.rV/ (such as *hōpara*), and WF3 /ˈrV.CV/ (such as *rapu*). The segmental environments include six /VrV/ sequences: /iri, ira, iro, ari, ara, aro/. These adjacent vowels function as point vowels in the Māori vowel space. The mid-back vowel /o/ was used in place of /u/, as the latter has become increasingly fronted and generally more rounded [16]. All target words analysed are summarised in Table 1.

3.2. Recording process and setup

Recording sessions were run in a WhisperRoom Sound Isolation Enclosure (`https://whisperroom.com`), with speakers seated at a desk in front of a computer monitor. A Rode Lavalier lapel microphone and Roland OCTA-CAPTURE were used for audio capture, pre-amplification, and digitisation. Audio was captured at a sample rate of 44.1kHz and 16-bit bit depth. Depending on the speech rate of the speaker and the speed at which they completed the recording process, up to five repetitions of each sentence were recorded. As a warm up task, speakers read aloud two passages of text in Māori shown on the monitor. This was followed by the central speech task, where sentences were displayed on the monitor and recorded one at a time. Recording sessions lasted a maximum of one hour. Of the 11 speakers analysed in the present study, one completed three repetitions, four completed four repetitions, and the remainder completed five repetitions.

3.3. Data preparation & processing

All processing, analysis, and visualisation of data was undertaken using R (version 4.2.0) [17]. Recordings were converted into *EmuR* database format using the package *EmuR* (version 2.3.0) [18]. WebMAUS General was used for a first pass at automated phonetics segmentation of recordings (language: language independent (SAMPA), otherwise default settings) [19]–[21]. Boundaries were then hand-corrected where necessary in *EmuR*. Start and end boundaries for the /r/ consonant were placed based on reduced amplitude of the waveform envelope and reduction in formant energies in the spectrogram and were placed at the nearest zero-crossing. Formant trajectories were estimated using the `forest()` function in the *wrassp* package using default settings aside from gender (gender set to '`f`' indicating 'female', which sets effective window length to 12.5 ms and nominal F1 to 560 Hz) [22]. All formant estimations were hand-adjusted by the first author.

The behaviours of first and second formants are visualised in the F1/F2 acoustic vowel space. This approach has been used to analyse formant patterns of Spanish consonants in /VCV/ sequences [23]. A similar approach is undertaken here, where three points in each /VrV/ sequence are extracted. These are (i) an approximation of the acoustic target of the preceding vowel, (ii) the temporal midpoint of /r/, and (iii) an approximation of the acoustic target of the following vowel. An approximation of vowel targets was necessary as these were not hand-labelled. Based on an analysis of present-day young speakers from the MAONZE corpus, in which vowel targets are labelled for long and short monophthongs, it was determined that, on average, the vowel target occurred 40% to 50% through the vowel production. The exact value depended on speaker and the particular vowel, and a relative position of 45% was selected to approximate the target position.

The first and second formant values (F1 and F2) were analysed using a linear mixed-effect (LME) model using the *R* package *lme4* (v. 1.1-35.3) [24]. Fixed effects were preceding vowel (either /i/ or /a/), following vowel (either /i/, /a/ or /o/), word form (WF1, WF2, or WF3), and phrase environment (P1 or P2). Separate models were constructed for F1 and F2 with model fitting informed by the `step()` function in the *lmerTest* package (v. 3.1-3) [25]. To determine significance of fixed effects, and their interactions, likelihood ratio tests (two-way ANOVA) were used. The *emmeans* package (v. 1.10.2) [26] was used for post-hoc pairwise comparisons. Speaker was added as a random effect in both models.

4. Results

First and second formant data was extracted from 607 tokens of /r/ (summarised in Table 1). The production study elicited an equal number of all segmental environments and word forms, however the proportion of tokens with formant energy differed across these.

Table 1: *Target words analysed the present study. Note WF indicates 'word form'.*

VrV	WF1		WF2		WF3		#
	word	#	word	#	word	#	
/iri/	piri	28	tāpiri	15	ripo	26	69
/ira/	hira	24	pākira	10	rapi	50	84
/iro/	piro	36	kōpiro	17	ropi	44	97
/ari/	pari	54	tōkari	34	ripa	54	142
/ara/	para	42	hōpara	15	rapu	55	112
/aro/	paro	37	tākaro	15	rotu	51	103
#	WF1	221	WF2	106	WF3	280	607

The distribution of /r/ targets in the F1/F2 space is illustrated in Figure 1, with colours differentiating the segmental environment that each /r/ is drawn from. The acoustic target approximations of the preceding and following vowels are also visualised in black, with letter label placement indicating their mean F1 and F2 values. The ellipses around vowels and /r/ targets were computed using `stat_ellipse()` in *ggplot2* (assumed t-distribution with 95% confidence level) [27]. As can be seen in the plot, there is sizeable area in the F1/F2 space in which the /r/ target occurs, and that this varies with segmental context. F1 in /r/ tends to be lowest in the when preceded by the high-front vowel /i/ (those points in dark blue, light blue, and pink) and higher when preceded by the open vowel /a/ (points in brown . When the /VrV/ sequence is a combination of the /a/ and a higher vowel, the F1 distribution is more spread across the vowel space.

We observe that F2 tends to be higher when /r/ is preceded by the high-front vowel /i/ when compared to those sequences which have only /a/ or /o/ preceding/following. There is some overlap between most segmental environments but there generally appear to be differences in all F2 distributions, save for /ari/ and /iro/ (shown in brown and pink, respectively).

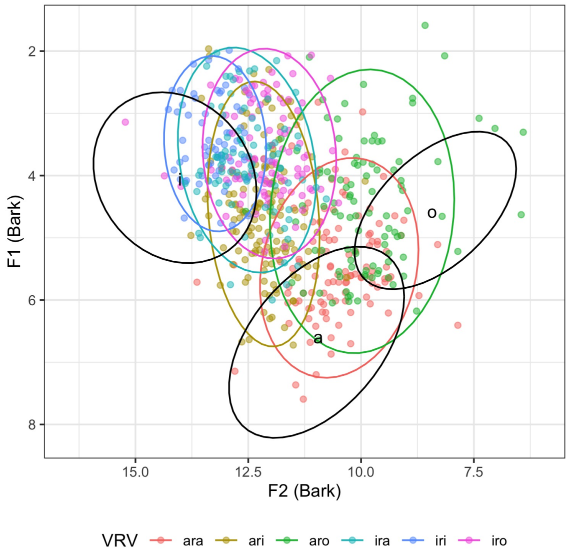


Figure 1: *F1 and F2 values (Bark-scaled) of all /r/ tokens.*

4.1. Modelling of the first formant (F1)

We used likelihood ratio tests to examine the main effects of the linear mixed-effect models. We found phrase context (described in Section 3.1) affected F1 ($\chi^2(1) = 7.41, p = 0.0065$), however post-hoc pairwise comparisons indicated there were no significant differences between phrase contexts. There were significant interactions between preceding and following vowel environment ($\chi^2(2) = 18.09, p = 0.0001$), and preceding vowel environment and word form ($\chi^2(2) = 27.05, p < 0.0001$).

Figure 2 shows the linear predictions of F1 for all levels of preceding and following vowel environment. The dots in this plot indicate the estimated marginal mean in each context, and error bars indicate the 95% confidence intervals. Both preceding and following vowel clearly influence F1. Based on these confidence intervals, preceding /i/ results in an /r/ F1 value between 3.18 and 4.03 Bark. When preceded by /a/, this value increases, ranging between 4.18 and 5.64 Bark. Post-hoc pairwise contrasts of following vowel were not significant when the preceding vowel was /i/. When the preceding vowel was /a/ and the following vowel was /i/ or /o/, pairwise contrasts with following /a/ were significant. There was no significant difference between F1 of /r/ in /ari/ and aro/. Pairwise contrasts are summarised in Table 2. The general trend we observe is that F1 of /r/ is clearly influenced by the preceding vowel, with preceding /i/ resulting in lower F1 than preceding /a/. In addition, we observe F1 in /r/ tends to be more susceptible to change due to following vowel when preceded by /a/ compared to /i/.

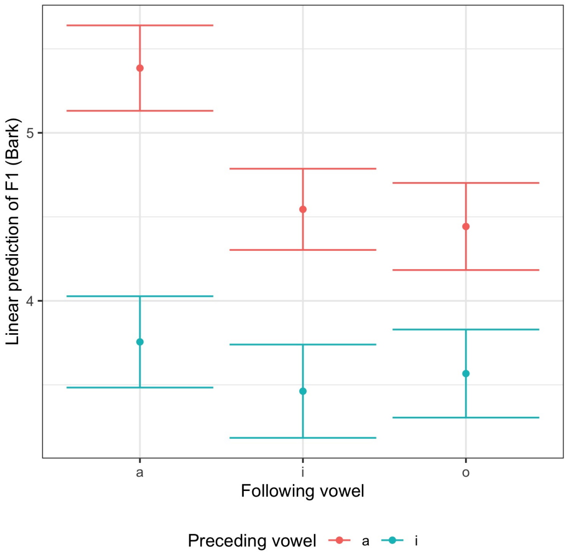


Figure 2: *Predictions of /r/ F1 for all vowel environments.*

Table 2: *Pairwise contrasts of following vowel (F1)*

Pre.	Contrast	Est. (SE)	df	t.ratio	p.value
a	a - i	0.84 (0.11)	592	7.83	$< .0001$
	a - o	0.94 (0.12)	589	8.23	$< .0001$

4.2. Modelling of the second formant (F2)

For F2, likelihood ratio tests found significant interactions between preceding and following vowel ($\chi^2(2) = 54.68, p < 0.0001$), preceding vowel and word form ($\chi^2(2) = 14.57, p = 0.0007$), and preceding vowel and phrase ($\chi^2(1) = 8.04, p = 0.0005$). Post-hoc pairwise comparisons of the latter indicated there were no significant differences between phrase contexts.

Linear predictions of F2 for all levels of preceding and following vowel are shown in Figure 3. Similar to Figure 2, this plot shows the estimated marginal mean and 95% confidence intervals. Similar to the case for F1, both preceding and following vowel environment influence F2. The general trend is for F2 to be lower when then preceding vowel is /a/ with the notable exception of /ari/, where /r/ is followed by the high-front vowel. In this case, F2 is increased to 12.16 Bark, overlapping almost totally with the range of values expected for /iro/ sequences.

For F2, pairwise contrasts of all following vowel combination were significant for both preceding /a/ and /i/. These are

Table of Contents for this manuscript

summarised in Table 3. The impact of following vowel on F2 appears to follow the same behaviour for preceding /i/ and /a/: F2 lowers when /r/ is followed by /a/ or /o/ (compared to /i/), and these changes are greater when /r/ is preceded by /a/.

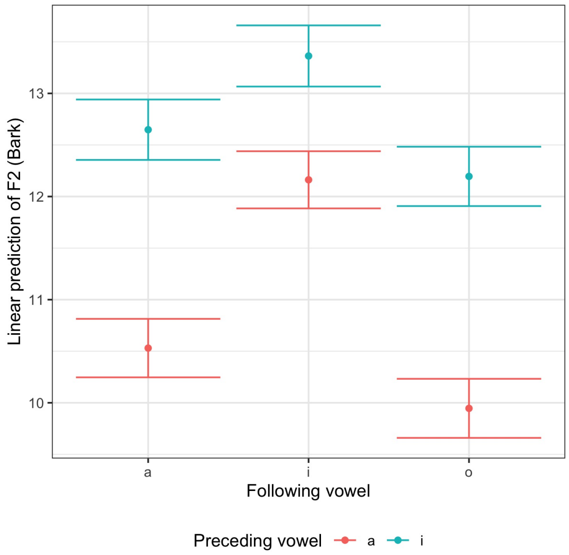


Figure 3: *Predictions of /r/ F2 for all vowel environments.*

Table 3: *Pairwise contrasts of following vowel (F2)*

Pre.	Contrast	Est. (SE)	df	t.ratio	p.value
a	a - i	-1.63 (0.09)	588	-18.77	< .0001
	a - o	0.58 (0.10)	586	6.30	< .0001
	i - o	2.22 (0.12)	587	24.95	< .0001
i	a - i	-0.72 (0.12)	586	-6.40	< .0001
	a - o	0.45 (0.10)	586	4.44	< .0001
	i - o	1.17 (0.11)	586	10.88	< .0001

As noted above, likelihood ratio tests found a significant interaction between preceding vowel and word form. Due to space constraints we do not include the full results of this analysis here. To summarise, post-hoc pairwise comparisons revealed that comparisons were only significant when the preceding vowel was /a/. This indicated that, when preceded by /a/, the F1 of /r/ in WF1 was higher than those in WF2 or WF3. The vowel preceding /r/ is stressed in WF1 and unstressed in WF2 and WF3, which may be the cause of this raising of F1. The pairwise comparisons of word form for F2 did not follow a coherent trend, with F2 of /r/ in WF1 and WF3 being higher than that in WF2.

5. Discussion

In this study we have investigated the impact of segmental environment, word form, and phrasal environment on the first and second formant behaviours of a subset of Māori /r/ realisations. These variants of the consonant have formant energy throughout their production, indicating there has not been complete closure during articulation. This realisation of the consonant has been observed to be the most common in previous work [14].

The present study has shown segmental context significantly impacts the F1 and F2 values of the consonant, providing some preliminary insight into coarticulatory changes in /r/. We can infer there are changes in constriction location (due to observed changes in F2) and of degree of constriction (due to F1). Such changes have been attested in other languages: vowel-dependent changes to closure place have been observed for taps in Catalan [28]. When preceded by the high-front vowel /i/, the F1 value at /r/ target did not change significantly, regardless of the following vowel. In the same context, F2 values did differ significantly depending on following vowel, but the influence was smaller compared to the preceding /a/ context. We may conclude then that /i/ is more restrictive than /a/ in this context, restraining /r/ F1 and F2 values to a range similar to that of the high-front vowel, although this could be natural consequence of vocal tract constraints.

As well as understanding variation in F1 and F2 behaviours of /r/, we were interested in visualising /r/ in the F1/F2 acoustic space along with targets of preceding and following vowels. This approach holds appeal as it places the unknown quantity (here, features of /r/) in a more familiar surrounding. Proctor used this approach to explore the acoustic targets of liquids in Spanish, identifying variation within and across the liquids [23]. It appears that the variation observed for the single rhotic of Māori may be greater than that observed in his study. It is possible that the variation observed here is due to the smaller consonant inventory of Māori: as there is no phoneme with which /r/ can be easily confused, more variation may be permitted. To explore this, comparison with other consonant phonemes in Māori or other languages should be considered.

This preliminary study reveals there are indeed distinctions in the F1/F2 behaviours of /r/, however we only considered these at a single point in time. The logical next step is to assess how the dynamic F1/F2 behaviours of such variants of Māori /r/ vary across segmental and word environments. This would provide more insight into the articulation of /r/, and would likely more meaningful than comparing what are effectively arbitrary points in time. We recommend further work consider these dynamic F1/F2 behaviours, as well articulatory analyses to complement such acoustic investigations.

6. Conclusions

This paper presents a preliminary investigation into F1 and F2 behaviours of Māori /r/ at a single point in time in its production. The present analysis considers only variants of the consonant with consistent formant energy during production. We consider the consonant in various segmental, word, and phrasal environments, and results suggest segmental environment impacts both F1 and F2 for all combinations of preceding and following vowels investigated. Word form also influenced F1 and F2, although only in some segmental environments. Phrasal environment was found to have no impact on either measure.

The observed influence of segmental environment on /r/ F1 and F2 indicates there is flexibility in the production of the consonant. The range of F1 and F2 values in /r/ is generally much more restricted when preceded (or followed) by /i/, suggesting there may be coarticulation processes at work. Further research is required to better understand this coarticulation and the interaction between Māori /r/ and its surrounding phones.

The findings of this study are limited as they only consider formant behaviours of Māori /r/ at a single point in time. We suggest future work consider dynamics of these formants in /r/ and adjacent vowels.

7. Acknowledgements

We thank the reviewers for their insightful comments. We also express our gratitude to the kaikōrero (speakers) who took part in this study.

8. References

[1] Statistics New Zealand, *2018 Census totals by topic*. 2018.

[2] B. Biggs, *Let's learn Maori: A guide to the study of the Maori language.* 1st edition. Wellington, New Zealand: A.H. & A.W. Reed., 1969.

[3] L. Thompson, C. I. Watson, H. Charters, R. Harlow, P. Keegan, J. King, and M. Maclagan, "An experiment in mita-reading: investigating perception of rhythmic prominence in the Māori language," in *Proceedings of SST 2010*, Melbourne, Australia, 14-16 December 2010.

[4] L. Thompson, C. I. Watson, R. Harlow, M. Maclagan, H. Charters, J. King, and P. Keegan, "Adventures in mita-reading: examing stress 'rules' and perception of prosodic prominence in the Māori language," in *Proceedings of ICPhS 2011*, Hong Kong, 17-21 August 2011.

[5] W. Bauer, *Maori*, 1st edition. London, UK: Routledge, 1993.

[6] B. Biggs, "The Structure of New Zealand Maaori," *Anthropological Linguistics*, volume 3, number 3, 1961.

[7] R. Harlow, *Māori: A linguistic introduction*. Cambridge, UK: Cambridge University Press, 2007.

[8] P. W. Hohepa, "A profile generative grammar of Maori," Ph.D. dissertation, Indiana University, 1965.

[9] D. B. Wilson, "A study of spoken Māori in Awarua (Northland)," M.S. thesis, University of Auckland, 1991.

[10] J. King, M. Maclagan, R. Harlow, P. Keegan, and C. Watson, "The MAONZE Corpus: Establishing a corpus of Māori speech," *New Zealand Studies in Applied Linguistics*, volume 16, number 2, pages 1–16, 2010.

[11] J. King, M. Maclagan, R. Harlow, P. Keegan, and C. Watson, "The MAONZE Corpus: transcribing and analysing Māori speech," *New Zealand Studies in Applied Linguistics*, volume 17, number 1, 2011.

[12] M. Maclagan and J. King, "A Note on the Realisation of /r/ in the Word Maori," *New Zealand English Journal*, volume 18, pages 35–39, 2004.

[13] M. Maclagan, T. Macrae, and J. Wilson Black, "The pronunciation of the word "Māori"," *Te Reo. Journal of the Linguistic Society of New Zealand*, volume 66, number 2, pages 130–153, 2024.

[14] I. Shields, C. Watson, and P. Keegan, "Preliminary analysis of /r/ acoustics and features in three Māori speakers," in *Proceedings of SST 2022*, Canberra, Australia, 13-16 December 2022.

[15] I. Shields, C. Watson, and P. Keegan, "Ngā āhuatanga o te /r/ o te reo Māori: preliminary investigations into the acoustics of Māori /r/," *Te Reo. Journal of the Linguistic Society of New Zealand*, volume 66, number 2, pages 105–131, 2024.

[16] M. Maclagan, C. I. Watson, R. Harlow, J. King, and P. Keegan, "/u/ fronting and /t/ aspiration in Māori and New Zealand English," *Language Variation and Change*, volume 21, number 2, pages 175–192, 2009.

[17] R Core Team, *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2024. [Online]. Available: https://www.R-project.org/.

[18] M. Jochim, R. Winkelmann, K. Jaensch, S. Cassidy, and J. Harrington, *emuR: Main Package of the EMU Speech Database Management System*, R package version 2.5.0, 2024. [Online]. Available: https://CRAN.R-project.org/package=emuR.

[19] F. Schiel and J. J. Ohala, "Automatic Phonetic Transcription of Non-Prompted Speech," in *Proceedings of ICPhS 1999*, 10.5282/ubm/epub.13682, San Francisco, CA, USA, 1-7 August 2011, pages 607–610.

[20] F. Schiel, "A Statistical Model for Predicting Pronunciation," in *Proceedings of ICPhS 2015*, Glasgow, UK, 10-14 August 2011.

[21] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, volume 45, pages 326–347, 2017.

[22] R. Winkelmann, L. Bombien, M. Scheffers, and M. Jochim, *wrassp: Interface to the 'ASSP' Library*, R package version 1.0.5, 2024. [Online]. Available: https://CRAN.R-project.org/package=wrassp.

[23] M. Proctor, "Gestural Characterization of a Phonological Class: the Liquids," Ph.D. dissertation, Yale University, 2009.

[24] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, volume 67, number 1, pages 1–48, 2015.

[25] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software*, volume 82, number 13, pages 1–26, 2017. DOI: 10.18637/jss.v082.i13.

[26] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, R package version 1.10.2, 2024. [Online]. Available: https://CRAN.R-project.org/package=emmeans.

[27] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016, ISBN: 978-3-319-24277-4. [Online]. Available: https://ggplot2.tidyverse.org.

[28] D. Recasens and M. D. Pallarès, "A study of /J/ and /r/ in the light of the "DAC" coarticulation model," *Journal of Phonetics*, volume 27, number 2, pages 143–169, 1999. DOI: https://doi.org/10.1006/jpho.1999.0092.

Sonority Patterns in Lelepa Onset Clusters

Chao Sun, Rosey Billington

Australian National University

Chao.Sun@alumni.anu.edu.au, rosey.billington@anu.edu.au

Abstract

Lelepa, an Oceanic language of central Vanuatu, has complex syllable structures with onsets of up to three consonants and codas of up to two, which is rare among Oceanic languages. This study examines two-consonant onset clusters using natural speech data, showing a preference for certain clusters and frequent violations of the sonority sequencing principle. Clusters with smaller sonority distances are preferred, which is uncommon crosslinguistically. These findings enhance our understanding of Lelepa's phonotactics and contribute to the broader typology of phonotactic constraints.

Index Terms: syllable structure, consonant cluster, phonotactics, Oceanic

1. Introduction

1.1. Lelepa language

Lelepa[1] is part of the Southern Oceanic linkage, closely related to languages like Nguna, Nafsan, and Eton. It is spoken by around 500 people on Lelepa and Efate islands in central Vanuatu. Lelepa is notable for its complex syllable structure, which is uncommon among Oceanic languages [1]. Previous linguistic description highlights that among all the complex syllable structures observed in Lelepa, two-consonant clusters in the onset position exhibit the most complexity, including some combinations that are uncommon crosslinguistically [2]. In the context of increasing interest in phonotactic typology, further investigation into these consonant clusters can enhance our understanding of structures in Lelepa and syllable typology more broadly.

1.2. Consonant sequencing within syllables

Many languages allow for complex syllable structures such as CVCC and CCV, although CV is widely considered the most basic and prevalent syllable structure [3, 4]. Generally, the complexity of syllables is conceptualised in relation to the overall structures they exhibit, i.e. the more consonants allowed in the onset or coda position of the syllable, the more complex the syllable is; complexity also relates to the specific segments that can occur within the onset, nucleus and coda, and in what combinations [5]. Thus, languages typically exhibit varying restrictions regarding the types of consonants occurring at syllable margins and their relative positions within a complex syllable. One of the most well-known restrictions is the Sonority Sequencing Principle (SSP).

Sonority is a concept widely used in phonological theory to explain the distribution of segments within the syllable (e.g. [6, 7]). It is generally understood to refer to the loudness or perceptual prominence of speech sounds, which correlates with the degree of constriction in the vocal tract and thus is closely associated with the manner of articulation [6]. Numerous studies have explored sonority and attempted to rank consonant categories based on their sonority levels. One commonly cited five-scale version [7] ranks segments as in (1):

$$\begin{aligned}&\mathit{Glides}\ (5) > \mathit{Liquids}\ (4) > \mathit{Nasals}\ (3) > \\ &\mathit{Fricatives}\ (2) > \mathit{Stops}\ (1)\end{aligned} \tag{1}$$

The SSP is a phonological principle that was first proposed in the late 19th century [8]. The SSP aims to account for reported crosslinguistic tendencies in the organisation of consonants within a cluster based on their sonority levels. A large-scale 1960s study of consonant clusters at the beginnings and ends of words in 104 languages highlights the role of sonority in predicting the frequency of allowed consonant clusters across different languages [9]. According to the SSP, an ideal well-formed sonority pattern exhibits a rising sonority from the left edge (the first consonant in the onset cluster) towards the nucleus, followed by a fall in sonority from the nucleus to the right edge (the last consonant in the coda cluster) [7, 9]. For example, the monosyllabic English word 'crisp' /kɹɪsp/ is considered to have the ideal sonority pattern according to the SSP. Within the onset position, which is the focus of the present study, a sonority rising slope, such as /plV/, is considered preferred, and a sonority falling slope, such as /lpV/, is not. A sonority plateau, such as /ptV/, is usually language-dependent and can be interpreted as SSP-followed or SSP-violated depending on the analysis (e.g. [6, 9]). Previous studies on sonority also note that in addition to a general preference for rising sonority, onset clusters with larger difference in sonority are preferred over clusters with smaller differences in sonority [7, 9]. For example, /plV/ should be preferred over /pnV/ according to the scale in (1).

However, while the sonority account has been influential in explaining typological tendencies in syllable structures, exceptions to the principle are not uncommon. Such instances can be found within languages like Russian, where the onset cluster /lb/ is relatively common, despite violating the SSP by exhibiting sonority reversal [10]. Another case of violation of the sonority sequencing principle is observed in the Georgian language, where sonority reversal in the onset position occurs in clusters such as /rk/ and /md/ [11]. Additionally, in English, the prevalence of onset clusters starting with the alveolar fricative /s/, which goes against the expected sonority pattern, cannot be accounted for by the sonority sequencing principle. Examples such as /st/ in the word 'stress' and /sp/ in the word 'spell' are quite common [6]. A recent large-scale study of lexical data for 496 languages across 58 language families finds that over half of the languages in the database permit clusters violating the SSP either in word-initial or word-final positions [12], indicating that violations of the SSP are not rare crosslinguistically.

These observations suggest that while sonority provides a useful framework, other phonetic, phonological, and language-

[1] ISO 639-3: lpa, Glottocode: lele1267

specific factors also influence cluster patterns in different languages. They also highlight another issue: current research on the SSP and phonotactic typology has predominantly focused on major, well-studied languages and relied primarily on text-based data such as lexicons or written corpora, often without syllabification, limiting analyses to word-initial and word-final clusters. Therefore, there is a need for closer investigation into a more diverse range of languages, including understudied and minority languages, as well as different data types such as spontaneous speech corpora. This approach can lead to a more comprehensive understanding of consonant cluster patterns.

1.3. Syllable structure in Lelepa

Table 1. *Lelepa consonant inventory (based on [2])*

	Bilabial	Labiodental	Alveolar	Palatal	Velar	Labial-velar
Plosive	p		t		k	k͡pʷ
Nasal	m		n		ŋ	ŋ͡mʷ
Fricative		f	s			
Trill			r			
Lateral			l			
Glide				j		w

The phoneme inventory of Lelepa includes 14 consonants as in Table 1, with no voicing distinctions among obstruents, which are all phonemically voiceless. In the grammatical description of Lelepa [2], a schema for the syllable structure is introduced, showing the maximum number of consonants allowed in the onset position as three, and two in the coda position. Ten syllable structures are identified in Lelepa, from simple CV to rare forms like CCVCC and CCCVC (see more detail in Table 4, Section 3.4). On this basis, in a recent crosslinguistic study of syllable structures, Lelepa has been classified as exhibiting 'complex' syllable structure [1]. Some uncommon consonant clusters are observed in the language, such as /ntV/, /rkV/ and /fsrV/, which do not follow the ideal form of the SSP [2].

Previous study observes that the presence of complex clusters in Lelepa relates to ongoing vowel deletion processes within the language [2], compared to more conservative related languages and the ancestral language, Proto-Oceanic. Some neighboring languages, such as Nafsan, have also undergone similar processes, seemingly somewhat earlier than in Lelepa [13, 14]. Table 2 shows two examples of the vowel deletion observed in Lelepa and Nafsan compared with Nguna, which is more conservative. Both the /pr/ and /mt/ onset clusters in the words /prau/ 'long' and /mtak/ 'afraid' arise due to the deletion of the word-medial vowel /a/ in the words /parau/ and /mataku/. While the cluster /pr/ is considered ideal according to the SSP, the cluster /mt/ is ill-formed. The presence of consonant clusters like /mt/, which may be dispreferred according to the SSP, has made the syllable structure of the Lelepa language more complex.

Table 2. *Words realised with vowel deletion in Lelepa [2] and Nafsan [13] compared with Nguna [15]*

Lelepa	Nafsan	Nguna	Gloss
prau	pram	parau	'long'
mtak	mtak	mataku	'afraid'

2. Research questions

Given that the complexity of consonant sequencing in Lelepa primarily revolves around the two-consonant onset clusters, this study builds on previous descriptive research [2] to undertake a quantitative investigation of Lelepa onsets based on natural speech data. The following questions are addressed:
1. What is the frequency of different consonant combinations in two-consonant onset clusters?
2. To what extent does the Sonority Sequencing Principle account for the onset patterns in Lelepa?

3. Method

3.1. Corpus and data

This study uses archived spoken language data, collected during fieldwork (2007-2012) in the Lelepa-speaking villages Natap̃ao and Mangaliliu and available as an open-access collection [2, 16]. The corpus includes approximately 100 items, totaling 13 hours of recordings. Each item includes audio or video recordings with time-aligned transcriptions in ELAN. Most of the annotation files from the archive include transcription, interlinear glossing, as well as a English freetranslation of the entire sentence.

For this study, a subset of the corpus comprising monologic narratives was selected. The subset was chosen based on genre, length, and recording quality. For genre, a variety of different types were selected to maximize the range of content covered, such as procedural descriptions, folktales, historical stories, and personal life stories. For length, samples of 2-5 minutes were chosen to allow for the analysis of a greater number of samples. Recording quality was assessed based on clarity, absence of ambient noise, and the voice quality of the speakers to ensure that the individual speech sounds could be segmented accurately. The dataset comprises 67 minutes and 48 seconds of recordings from 13 male speakers and 8 female speakers (for further details see [17]).

3.2. Data processing and segmentation

The ELAN [18] files for the chosen subset were first manually verified, then exported as Praat textgrids and subsequently separated into utterance-level files using a script in Praat [19]. Semi-automatic phone segmentation and labelling were performed using the online MAUS tools [20, 21], with manual correction of phone boundaries to ensure accuracy. To address the research questions in this study, additional annotations were then added to mark syllable structures.

3.3. Syllabification

In the case of Lelepa, where words can be polymorphemic and polysyllabic, consistent syllabification, particularly of word-medial consonant sequences like ...VCCCV... is crucial for analyzing Lelepa's syllable structure from actual speech data. The approach to syllabification used in this study follows the syllabification previously described for Lelepa word-medial consonant sequences [2]. According to this approach, in word-medial heterosyllabic consonant clusters, which can consist of up to three consonants, the first consonant is considered the coda of the preceding syllable, while the remaining consonants form the onset of the next syllable. This means that the sequence ...VCCCV... is syllabified as ...VC.CCV.... A summary of the approach is shown in (2), which demonstrates the hierarchy of roles for word-medial consonants in polysyllabic words:

$$\begin{aligned} &singleton\ onset > singleton\ coda \\ &> onset\ cluster > coda\ cluster \end{aligned} \qquad (2)$$

For example, consider the syllabification for the words /ar=msoun/ '3DU.R=want' and /e=salpnot/ '3SG.R=float.come'. In the case of the word /ar=msoun/, there is a word-medial three consonant sequence which is /rms/, and according to the rules mentioned above, the first consonant /r/ functions as a coda consonant following the nucleus vowel /a/, while the others /ms/ form the onset cluster for the next syllable /msoun/. It is the same with the second word; /pn/ forms the onset cluster for the syllable /pnot/. The syllabification of these is summarised in Table 3.

Table 3. *Example of syllabification*

ar=msoun	VCCCVV	→	[ar.msoun]	VC.CCVVC
e=salpnot	VCVCCCVC	→	[e.sal.pnot]	V.CVC.CCVC

3.4. Database and analysis

After the syllabification process using Praat, the Praat textgrid files were converted into a hierarchical database using the EMU Speech Database Management System, accessible through R with the emuR package [22]. This allowed for filtering and extracting syllable structure information, resulting in a dataset of 10,971 syllables. These occur in words of up to seven syllables, but there is a preference for words of three syllables or fewer. 42.8% of words are disyllabic, 25.3% are trisyllabic and 21.8% are monosyllabic. Table 4 shows the frequency of syllable structures found in the dataset by token and type, where 'tokens' indicates the total occurrences of each combination, and 'types' represents the number of unique words containing these clusters. The frequency of different syllable structures is almost identical for both tokens and types. The most frequent syllable structure is CV, with 6092 tokens. The next most frequent structures, CVC, V, and VC, do not allow consonant clusters. Structures with two-consonant onset clusters, namely CCV, CCVC and CCVCC, collectively account for 596 tokens.

Table 4. *Frequency of syllable structures in the dataset*

Syllable structure	Tokens	Percentage (%)	Types	Percentage (%)
CV	6092	55.53	1245	43.27
CVC	2834	25.83	865	30.07
V	887	8.08	325	11.30
VC	526	4.79	208	7.23
CCV	423	3.86	133	4.62
CCVC	159	1.45	80	2.78
CVCC	25	0.23	15	0.52
CCVCC	14	0.13	2	0.07
CCCVC	6	0.05	3	0.10
CCCV	5	0.05	1	0.03

Each consonant in the two-consonant onset clusters was assigned a value according to the sonority scale in (1). The sonority slope and sonority distance can be calculated by subtracting the sonority value of the first consonant in the cluster from that of the second consonant. For example, for a cluster like /tl/, the sonority distance is $4 - 1 = +3$, which indicates a sonority rising slope. For a cluster like /sf/, the sonority distance is $2 - 2 = 0$, which indicates a sonority plateau, and for a cluster like /mt/, the sonority distance is $1 - 3 = -2$, which indicates a sonority falling slope.

4. Results

4.1. Sonority slope

Table 5. *Onset two-consonant clusters grouped by manner of articulation. The frequency of 'types' is the number to the left, and the frequency of 'tokens' is the number in brackets.*

		Consonant 2				
		PLO	FRI	NAS	LIQ	GLI
Consonant 1	PLO	11 (20)	7 (8)	3 (5)	26 (38)	3 (3)
	FRI	19 (106)	3 (12)	2 (2)	21 (67)	1 (1)
	NAS	25 (72)	14 (21)	27 (146)	27 (50)	3 (27)
	LIQ	9 (11)	-	5 (7)	-	-
	GLI	-	-	-	-	-

In the dataset, 50 unique consonant combinations were identified in the 596 syllable tokens with two-consonant onset clusters, occurring within 206 different words. These clusters are categorized by their manner of articulation and are presented in Table 5. In this table, 'Consonant 1' refers to the consonant in the syllable-initial position, while 'Consonant 2' refers to the consonant closer to the nucleus. Each combination shows the count in tokens and types as mentioned earlier: the number on the left represents the number of unique words containing these clusters (types), and the number in parentheses indicates the total occurrences (tokens). For example, the entry '7 (8)' in the first row signifies that the plosive-fricative onset cluster appears eight times across seven different words. Table 5 shows that a range of consonant combinations are possible, and sequences starting with a nasal followed by either another nasal or a liquid are more common, each observed 27 times, followed by plosive-liquid clusters (26 times) and nasal-plosive clusters (25 times).

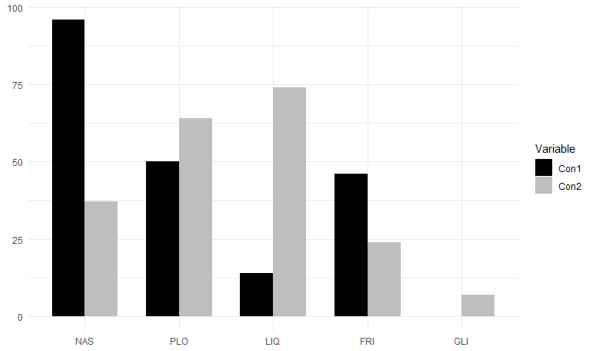


Figure 1: *Consonant distribution of onset two-consonant clusters (types)*

From Table 5, we can compare the frequency of each group of consonants in the Consonant 1 or Consonant 2 position in two-consonant onset clusters. Figure 1 shows this comparison more clearly. According to Figure 1, nasals are the most frequent consonants in the Consonant 1 position, occurring in 96 unique words, while liquids are the most frequent in the Consonant 2 position, occurring in 74 words. Glides do not occur in the Consonant 1 position.

Table 5 also indicates the change in sonority within the clusters. The gray regions represent a sonority plateau, where the sonority change is 0 as introduced in Section 3.4, such as in a nasal-nasal sequence. These gray regions also divide the table into three parts: the regions to the upper right indicate a sonority rise, such as in a plosive-fricative sequence, and the regions to the lower left indicate a sonority fall, such as in a nasal-plosive sequence. This demonstrates that all three patterns of sonority slopes can be observed. Figure 2 further illustrates the frequency of different sonority slopes in two-consonant onset clusters. Most clusters exhibit a sonority rising slope (45.1%), and while 19.9% of clusters exhibit a plateau in the onset position, 35.0% of the clusters show a sonority falling slope, which violates the Sonority Sequencing Principle.

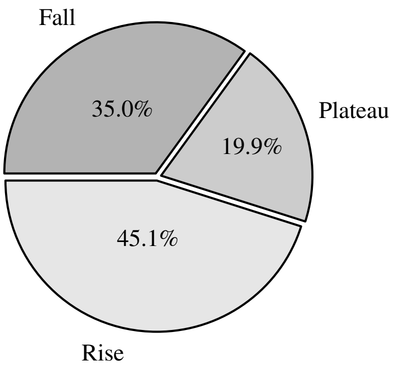


Figure 2: *Sonority slopes of onset two-consonant clusters (types)*

4.2. Sonority distance

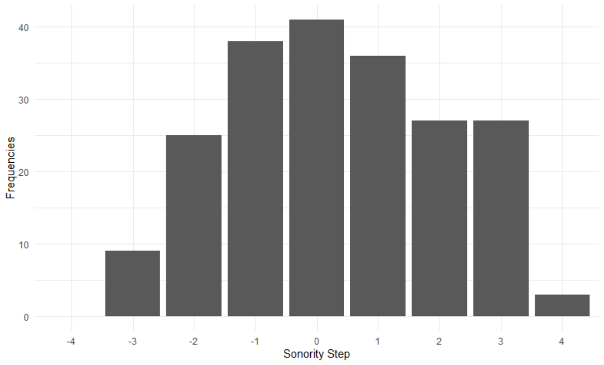


Figure 3: *Sonority distances of onset two-consonant clusters (types)*

Based on the five-scale sonority rank mentioned in (1), Figure 3 shows the frequency of different sonority distances for the two-consonant onset clusters. The sonority distance ranges from -3 (e.g., liquid-plosive) to +4 (e.g., plosive-glide). The most common sonority distance is 0, such as in nasal-nasal clusters, occurring 41 times. This is followed by a sonority distance of -1, occurring 38 times, and +1, occurring 36 times. This shows that onset clusters with small sonority distances are preferred. Although sequences with a large sonority distance, such as +4, are found in the database, they are the least common, occurring only 3 times.

5. Discussion and conclusions

This study focuses on the sonority pattern of two-consonant onset clusters in Lelepa. The results, based on both syllable tokens and types, show that Lelepa allows for most consonants in both positions of the onset cluster, except that glides only occur in the Consonant 2 position. The frequency of clusters in Table 5 shows that the most common clusters are nasal-nasal clusters and nasal-liquid clusters, each accounting for 13.1%. The frequency of consonants in Figure 1 shows that the most common consonants in the Consonant 1 position of the cluster are the nasals, accounting for 46.6%, while the most common consonants in the Consonant 2 position are the liquids, accounting for 35.9%. The prevalence of initial nasals likely relates to the very common nominal marker /n(V)/ that occurs in Lelepa and other central Vanuatu languages (discussed further below) [23].

The data shows that, as expected, all three types of sonority slopes - rising, plateau, and falling - are allowed in the onset position in Lelepa. According to the Sonority Sequencing Principle (SSP) introduced in Section 1.2, clusters with a sonority rising slope are expected to be preferred in the onset position. However, the frequency of clusters in Figure 2 shows that less than half exhibit sonority rising slopes (45.1%). The rest exhibit either sonority plateaus (19.9%) or sonority falling slopes (35.0%), which would typically be considered a violation of the SSP. However, as mentioned in Section 1.2, there is increasing evidence that violations of the SSP are not rare crosslinguistically [12]. The present study lends support to these observations, and highlights the need for more intensive exploration of phonotactic typology with reference to factors beyond sonority.

The SSP also suggests that the frequency of different clusters can be related to the degree of sonority distance between the consonants within the cluster, with a greater sonority distance being preferred in the onset position. However, according to Figure 3, the most common clusters are those with 0 sonority distance, namely the sonority plateau clusters. This result does not align with the sonority distance preferences of the SSP. As mentioned earlier in Section 1.3, vowel deletion in Lelepa mainly drives the formation of consonant clusters. One possible explanation for the lack of large sonority distances in onset clusters is that the (relatively recent) deletion process allows a high degree of flexibility in which consonants can be adjacent to one another. Additional clues can be found in other languages of central Vanuatu, such as Nafsan [13] and Neve‘ei [24], which also permit consonant clusters violating the SSP and have also undergone medial vowel deletion processes. In all three languages, the clusters that violate the SSP often begin with the alveolar nasal /n/, which aligns with the frequency distribution of consonants in Figure 1, despite the crosslinguistic rarity of /n/ serving as the initial consonant in SSP-violating clusters [12].

The findings of this study confirm the complexity of Lelepa syllables, and suggest the need to consider language-specific factors and how they might interact with previously observed crosslinguistic tendencies, such as sonority sequencing. This study also highlights the need to continue expanding the range of languages analyzed in research on phonotactic typology, particularly focusing on segment sequencing in clusters. It would be beneficial for crosslinguistic research in this area to incorporate data from languages representing a broader range of language families and linguistic structures, with a greater focus on understudied languages.

6. Acknowledgements

We wish to acknowledge the Lelepa speakers who contributed to the corpus this study draws on, and Sébastien Lacrampe for providing access to the Lelepa materials via the ELAR archive.

7. References

[1] Easterday, S. *Highly Complex Syllable Structure: A Typological and Diachronic Study*. Berlin: Language Science Press, 2019.

[2] Lacrampe, S. “Lelepa: Topics in the grammar of a Vanuatu language,” Ph.D. Thesis, The Australian National University, Canberra, 2014.

[3] Jakobson, R. *Selected Writings I: Phonological Studies*, The Hague: Mouton, 1962.

[4] Zec, D. “The syllable,” in *The Cambridge Handbook of Phonology*, 2007, pp. 161–194.

[5] Maddieson, I. “Syllable structure,” in *The World Atlas of Language Structures Online (v2020.3)*. Zenodo, 2013. doi: 10.5281/zenodo.7385533.

[6] Blevins, J. “The syllable in phonological theory,” in *The Handbook of Phonological Theory*, 1st ed. Cambridge: Blackwell, 1995, pp. 206–244.

[7] Clements, G. N. “The role of the sonority cycle in core syllabification,” in *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, Cambridge: Cambridge University Press, 1990, pp. 283–333.

[8] Sievers, E. *Grundzüge der Phonetik: Zur Einführung in das Studium der Lautlehre der Indogermanischen Sprachen*. Leipzig: Breitkopf und Härtel, 1881.

[9] Greenberg, J. H. “Some generalizations concerning initial and final consonant sequences,” Linguistics, vol. 3, no. 18, pp. 5–34, 1965, doi: 10.1515/ling.1965.3.18.5.

[10] Pouplier, M., Marin, S., Hoole, P., and Kochetov, A. “Speech rate effects in Russian onset clusters are modulated by frequency, but not auditory cue robustness,” *Journal of Phonetics*, vol. 64, pp. 108–126, 2017. doi: 10.1016/j.wocn.2017.01.006.

[11] Crouch, C., Katsika, A., and Chitoran, I. “Sonority sequencing and its relationship to articulatory timing in Georgian,” *Journal of the International Phonetic Association*, vol. 53, no. 3, pp. 1–24, 2023. doi: 10.1017/S0025100323000026.

[12] Yin, R., van de Weijer, J., and Round, E. R. “Frequent violation of the Sonority Sequencing Principle in hundreds of languages: How often and by which sequences?,” *Linguistic Typology*, vol. 27, no. 2, pp. 381–403, 2023. doi: 10.1515/lingty-2022-0038.

[13] Billington, R., Thieberger, N., and Fletcher, J. “Nafsan,” *Journal of the International Phonetic Association*, vol. 53, no. 2, pp. 511–531, 2023. doi: 10.1017/S0025100321000177.

[14] Billington, R., Thieberger, N., and Fletcher, J. “Phonetic evidence for phonotactic change in Nafsan (South Efate),” *Italian Journal of Linguistics*, vol. 32, no. 1, pp. 125–150, 2022. doi: 10.26346/1120-2726-151.

[15] Schmidt, H. *Nguna Dictionary*, Neuendettelsau: Erlanger Verlag für Mission und Ökumene, 2023.

[16] Lacrampe, S. “Possession in Lelepa, a language of Central Vanuatu,” Master Thesis, University of the South Pacific, Suva, 2009.

[17] Sun, C. “An investigation of syllable structure in Lelepa,” Master Thesis, The Australian National University, Canberra, 2023.

[18] The Language Archive, *ELAN*. (version 6.5) Nijmegen: Max Planck Institute for Psycholinguistics, 2023. Accessed: May 20, 2023. [Computer program]. Available: https://archive.mpi.nl/tla/elan.

[19] Boersma P. and Weenink, D., *Praat: Doing Phonetics by Computer*. (version 6.3.10), 2023. Accessed: May 20, 2023. [Computer program]. Available: http://www.praat.org/

[20] Reichel, U. D. “PermA and Balloon: Tools for string alignment and text processing,” in *Proceedings of INTERSPEECH 2012*, ISCA, 2012, pp. 1874–1877.

[21] Kisler, T., Reichel, U. D., and Schiel, F. “Multilingual processing of speech via web services,” *Computer Speech & Language*, vol. 45, pp. 326–347, 2017. doi: 10.1016/j.csl.2017.01.005.

[22] Winkelmann, R., Harrington, J., and Jänsch, K. “EMU-SDMS: Advanced speech database management and analysis in R,” *Computer Speech & Language*, vol. 45, pp. 392–410, 2017. doi: 10.1016/j.csl.2017.01.002.

[23] Lynch, J. “Article accretion and article creation in Southern Oceanic,” *Oceanic Linguistics*, pp. 224–246, 2001. doi: 10.1353/ol.2001.0019.

[24] Musgrave, J. *A Grammar of Neve’ei, Vanuatu*. Canberra: Pacific Linguistics, 2007.

Extending ASR Systems Error Measurements: Reporting Lexical and Grammatical Errors

Simon Gonzalez, Jason Littlefield, Tao Hoang, Maria Kim, Tim Cawley, Jennifer Biggs

Defence Science and Technology Group

simon.gonzalez@defence.gov.au, jason.littlefield@defence.gov.au, tao.hoang@defence.gov.au, myung.kim@defence.gov.au, tim.cawley@defence.gov.au, jennifer.biggs@defence.gov.au

Abstract

We address the limitations of the current Automatic Speech Recognition evaluation metric Word Error Rate. While used for broad assessment, it lacks the granularity to discern errors in specific linguistic categories. We offer a metric based on parts of speech and grammatical categories. Using the Whisper ASR system on English, Japanese, and Spanish, within the CommonVoice 15, we analyse Grammatical and Lexical error rates. Results show that Grammatical words trigger less errors than Lexical words, and case markers combined with Lexical words in Japanese, trigger higher accuracy. Our approach enhances the explanatory power of error analysis in ASR system performance.

Index Terms: Automatic Speech Recognition, error metrics, parts of speech, lexical and grammatical evaluation

1. Introduction

Automatic Speech Recognition (ASR) technologies have undergone significant advancements [1][2] and the widespread adoption of ASR systems in various industries (e.g., Healthcare, Defence and Automotive) highlight the critical role of accurate evaluation to ensure their effectiveness, reliability, and user satisfaction. However, evaluating the performance of ASR systems remains a challenging task [3]. Traditional rate measurements, such as the widely used Word Error Rate (WER), offer valuable insights into overall system performance. WER is calculated as the ratio of the total number of errors – comprising substitutions, deletions, and insertions in the transcription output – to the number of words in the audio signal input to the ASR system [4]. But WER has been reported to have limitations [5]. The primary limitation of WER lies in treating all errors uniformly, without distinguishing between those fully detrimental to the meaning of the ground truth reference and those with closer semantic or syntactic relevance. Furthermore, WER cannot gauge the relative importance of specific words in the ground truth transcription, prompting the proposal of alternative metrics that account for semantics [6], entity recognition [7], and parts of speech [8]. Previous research also indicates that WER does not consistently correlate with human judgment on ASR system performance [3][9]. These findings underscore the need for linguistic metrics offering a more detailed understanding of errors, moving beyond the holistic view currently provided. As the field of ASR matures, there is a growing recognition that a more detailed analysis of errors is imperative for refining and advancing these technologies. One limitation of existing metrics is their inability to unveil the specific nature of errors. While systems may exhibit similar overall error rates, this metric fails to elucidate whether these errors target distinct linguistic categories. For example, two ASR systems may boast comparable WERs, yet one might prove more detrimental to Grammatical words, while the other, with a seemingly identical error rate, might manifest more errors in Lexical or content words. To address this deficiency, there is a pressing need to delve deeper into the intricacies of errors by examining linguistic categories, thereby shedding light on the distinct areas of vulnerability within ASR systems [10][11].

Recognizing the limitations of current approaches, we propose the integration of linguistic metrics in the evaluation of ASR systems. An in-depth analysis based on linguistic categories, including parts of speech and grammatical classifications, can help in understanding the complexities of errors. By categorizing errors according to linguistic attributes, valuable insights can be gained into the nature of errors and how they behave within the context of these systems. This linguistic perspective not only brings clarity to the nature of errors but also enhances the explanatory power of error analysis in ASR, providing a more comprehensive understanding of system performance.

The purpose of our current approach is to present methodology for analyzing and reporting errors in ASR systems. Taking a multilingual approach, we analyze errors in English, Japanese, and Spanish, leveraging the Whisper ASR system [12] on the CommonVoice 15 dataset [13] described in Section 3 below. Here, we propose analyzing errors based on word classes, grouped into two major categories: Grammatical words and Lexical words. Utilizing Parts of Speech tagging (POS) (described in Section 3.4.2), we measure Grammatical/Lexical errors, breaking down errors into two distinct values: one for Grammatical or function words and another for Lexical or content words. This approach aims to provide a more detailed and informative perspective on the performance of ASR systems, catering to the need for specific error analysis relevant for speech recognition technologies.

2. Previous work and current approach

There have been important advancements in methodologies for measuring ASR performance, including the adoption of various error metrics. Recent efforts within ASR evaluation have shifted focus towards metrics that go beyond word counts, such as *word embeddings* [14], *sentence embeddings* [15], and *semantic proximity* [16]. Taking inspiration from machine translation, where linguistic metrics have proven instrumental in refining translation quality, this paper extends the discourse to ASR. Works such as [17] successfully integrated linguistic attributes, like parts of speech, into translation evaluation, extending beyond WER and BLEU scores. They proposed the

Position Independent Error Rate (PER) across different parts of speech, estimating the contribution of each POS class to the overall word error rate. Although their work involved errors based on POS in two languages, English and Spanish, and compared them to human assessors, the remaining question is whether these findings can be generalized to other languages with different typological characteristics. Our goal is then to build from this by developing metrics within a single widely used ASR system, facilitating a comparison between the reference (**REF**) and the hypothesized text (**HYP**) produced by the ASR system across multiple languages with major typological differences.

We aim to improve upon the work of [8], who proposed a linguistic-based error metric, offering a finer-grained analysis of errors and discrepancies. We aim to do this in two ways. Firstly, we compare three languages, English, Japanese, and Spanish, all with different levels of inflections. Linguistic inflections are changes in the form of a word to mark distinctions such as tense, person, and number. For example, verb conjugations are a type of inflections and regular plurals in English. For our second improvement, we propose making distinctions between Grammatical and Lexical categories, grouping POS categories into these classes since their errors have different impacts on the **HYP** text. We make this distinction because Lexical errors directly contribute to the misunderstanding of the message and the incorrect interpretation of the text [18]. In contrast, Grammatical errors, while potentially leading to misunderstanding, have a less disruptive impact than Lexical errors. This distinction enables us to compare ASR errors not only at the overall level but also how they manifest concerning Lexical and Grammatical words. Such an approach contributes to the understanding of ASR errors, addressing the existing literature gap in the exploration of linguistic metrics, particularly in the context of evaluating errors based on Grammatical and Lexical categories. In summary, while existing metrics have laid a solid foundation for ASR evaluation, our work contributes by incorporating linguistic metrics inspired by both ASR advancements and successful methodologies in machine translation. We aim to fill the current gap in detailed error analysis within the ASR domain, adapting linguistic metrics to provide a more comprehensive understanding of errors.

3. Methodology

Developing performance metrics is a substantial undertaking. Based on [9] and [19], we developed our metrics to meet four criteria deemed crucial for a metric to possess. Firstly, it should reflect some level of human judgment, aiding in the identification of how much information is effectively communicated and how much is lost. Secondly, it must be straightforward to apply, which is a crucial feature when comparing across different ASR systems. Thirdly, it should be language-independent, which helps in comparing errors across languages from different typological classifications and does not favor one language structure over another. Finally, the metric should be easy to interpret from the outputs. In the development of the metrics presented here, we adhered to these principles to align with the practicality and real-world applicability prevalent in the field.

3.1. Languages chosen

The selection of languages was driven by both data availability and the authors' expertise, resulting in the choice of English, Japanese, and Spanish. These languages serve as robust testing grounds due to their shared characteristics and notable differences. Both English and Spanish belong to the Indo-European language family, and Japanese belongs to the Japonic language family [20]. They also exhibit divergences in their levels of inflection, a factor relevant to ASR system errors. Research has indicated that word classes with higher inflection are more prone to errors compared to those with less or no inflection [21]. For instance, the English article *the* remains uninflected, while its Spanish counterparts (feminine singular: “la”, masculine singular: “el”, feminine plural: “las,” masculine plural: “los”) carry gender and number inflections. Additionally, variations in inflection levels are evident in verb paradigms. While English may have six main forms (base, infinitive, past simple, past participle, gerund, and third person singular) [22], Japanese has 12 inflections [23], and Spanish can have 52 distinct forms reflecting person, number, tense, aspect, and mood [24]. These linguistic differences in inflection levels contribute to the richness of errors observed in ASR systems.

3.2. Speech datasets

This study utilized the Common Voice 15 dataset, a publicly available collection of multilingual and open voice data provided by the *Mozilla Common Voice Project* [13]. Designed for training and validating automatic speech recognition systems, the dataset encompasses a diverse range of voices and linguistic contexts. Table 1 below displays the characteristics of the datasets per language.

Table 1:.*Dataset descriptions for each language.*

Descriptors	**EN**	**JA**	**SP**
Total Number of Files	*16,386*	*4,978*	*15,796*
Total Duration	*26.9 hr*	*6.6 hr*	*26.8 hr*
Average File Duration	*5.9 sec*	*4.8 sec*	*6.11 sec*
Total Characters	*890K*	*105K*	*960K*
Total Words	*153K*	*55K*	*156K*
Unique Words	*21K*	*8K*	*23K*

The dataset encompasses contributions from a substantial number of speakers, providing a rich variety of linguistic and acoustic characteristics. In our analysis, we focused on a subset consisting of recordings from the test sets for the three languages. The dataset comprises over 16,000 sentences for English, approximately 5,000 for Japanese, and more than 15,000 sentences for Spanish. This offers a comprehensive sample of spoken language for evaluating ASR systems. The inclusion of a broad range of sentences and speakers enhances the robustness and generalizability of our findings, contributing to a more comprehensive understanding of the performance of the ASR system in diverse linguistic contexts. This includes variations in syntactic, semantic, and phonetic-phonological contexts.

3.3. Speech datasets

All our experiments were conducted using *OpenAI Whisper* [12]. Whisper comprises multilingual multitask models trained on 680,000 hours of labeled and curated speech data from diverse internet sources. In this experiment, we employed *Whisper-Tiny* (**T**), *Whisper-Medium* (**M**), *Whisper Large-v2* (**LV2**) and *Whisper Large-v3* (**LV3**). Comparing these four model sizes allows us to examine whether there are relevant accuracy gains across all language models.

3.4. Analysis

3.4.1. *Word Error Rate*

To assess the performance of the ASR system, we utilized the WER metric, a widely accepted measure for transcription accuracy assessment [25]. WER is computed by comparing the reference transcript (ground truth) with the output generated by the ASR system. The formula for WER is given by:

$$\text{WER} = (\text{S}+\text{D}+\text{I}) / \text{N}$$

Where, *S* represents the number of substitutions, *D* represents the number of deletions, *I* represents the number of insertions, and *N* is the total number of words in the reference transcript.

The analysis was conducted in R [26] using the outputs of Whisper. Our focus lies in ASR errors when comparing the reference text (**REF**) to the hypothesis text (**HYP**). SCLITE was employed for error calculation, identifying substitutions, insertions, and deletions per sentence. SCLITE, part of the NIST SCTK Scoring Toolkit, is a tool for scoring and evaluating speech recognition system output. It compares the **HYP** to the correct **REF**. Post-comparison statistics are gathered, and various reports can be generated to summarize recognition system performance.

3.4.2. *Parts of Speech and Lexical Tagging*

Linguistic tagging was conducted using the UDPIPE library [27] in R to enhance the textual analysis of transcribed speech data. UDPIPE, a state-of-the-art natural language processing (NLP) library, incorporates pre-trained models for various linguistic tasks. Specifically, we employed UDPIPE's pipeline for POS tagging. The tagging process consisted of three main steps.

Firstly, in text preprocessing, raw transcripts underwent preprocessing to eliminate artifacts or noise that might impact tagging accuracy (e.g. punctuation, case sensitivity, and text normalization), which allows for accurate comparisons between **REF** and ASR-generated transcripts. Secondly, during tokenization, preprocessed transcripts were tokenized into individual words or sub-word units using UDPIPE's tokenization module. The third step involved POS Tagging, where the POS tagging module assigned grammatical categories – such as nouns, verbs, adjectives – to each token in the transcripts. This information was crucial for understanding the syntactic structure of the spoken content.

3.4.3. *Linguistic Metrics Analysis*

We propose a metric that categorizes errors based on whether they occur in any of the two categories within a *Word Class*: GRAMMATICAL Words or LEXICAL Words. From the UDPIPE output, each POS was grouped into either the GRAMMATICAL group (ADP, AUX, CCONJ, DET, PART, PRON, SCONJ) or the LEXICAL Group (ADJ, ADV, NOUN, NUM, PROPN, VERB). From this, we calculated errors at the POS tagging in the **REF** and **HYP** texts, defined as POS_er, following the formula below:

$$\text{Pos\_er} = (\text{S}_{\text{POS}}+\text{D}_{\text{POS}}+\text{I}_{\text{POS}}) / \text{N}_{\text{POS}}$$

Word Class errors are then calculated for each group across the entire dataset per language and language model size: LEXICAL errors (LEX_er) and GRAMMATICAL errors (GRAM_er). To ensure fair comparisons, the analysis was conducted on sentences with matching number of words, i.e. when **REF** and **HYP** have the same number of words, avoiding penalization for incorrect pairs due to deletions and insertions. After comparing the accuracy metrics for POS, we then carried on further analysis explore the role of predictable linguistic patterns with the LEXICAL and GRAMMATICAL categories.

4. Results

Table 2 provides a summary of the experiment results, highlighting observable differences in performance across the **T**, **M**, **LV2** and **LV3** models for all evaluated languages. The **T** models consistently show the highest error rates (English = 23.7%; Japanese = 24.6%; Spanish = 23.5%), while the other models (**M**, **LV2** and **LV3)** demonstrate notably lower and more uniform WERs across all languages. Among these, the **LV3** model yields the most accurate results (English = 8.3%; Japanese = 5.7%; Spanish = 4%). It is evident that the **T** models show comparable WERs for the three languages, whereas the larger models exhibit higher accuracy, with Spanish being the most accurate and English the least accurate.

Table 2. *Breakdown of error rates results.*

Lang.	**Category**	**Language Model**			
		LV3	**LV2**	**M**	**T**
EN	WER	8.3%	8.9%	9.9%	23.7%
	POS_er	5.2%	5.5%	6.1%	17%
	LEX_er	11%	11.4%	11.9%	22.4%
	GRAM_er	4.6%	4.8%	5.3%	15.1%
JA	WER	5.7%	6.4%	7.5%	24.6%
	POS_er	1.7%	2%	2.4%	12.7%
	LEX_er	3.6%	3.9%	5.4%	22.9%
	GRAM_er	1.9%	2.7%	3.2%	13.7%
SP	WER	4%	4.9%	5.8%	23.5%
	POS_er	1.5%	1.9%	2.2%	10.9%
	LEX_er	4.2%	4.7%	5%	12.4%
	GRAM_er	1.6%	1.3%	3.4%	8.5%

When delving into the other metrics, a more detailed understanding emerges, shedding light on the categories to which ASR systems are more susceptible for errors. POS_er results demonstrate lower error rates in comparison to WER. This is notably more distinctive for Japanese and Spanish (English = 8.3%WER vs 5.2% POS_er; Japanese = 5.7% vs 1.7%; Spanish = 4% vs 1.5%). These results indicate that errors are more generalizable at the POS level, as compared to the word level. As such, this can help better our understanding of what types of errors can be consistently expected from ASR outputs, and in what morphological contexts.

A more in-depth analysis looked at those cases where the word form was incorrect (which counts to more WER) but it still had the same POS (which did not count as error for the POS_er). In case of inflectional languages, this difference can be observed when the **HYP** text has a singular form of a noun (e.g. *cat*), but the **REF** was the same word in the plural form (e.g. *cats*). The purpose was to see how much linguistic information is not captured if we stopped at the WER level. Figure 1 below shows a breakdown by language and model size for this experiment. It shows that Japanese and Spanish have more cases where errors are explained by inflectional differences between words (i.e. words are different but not their POS), as compared to English.

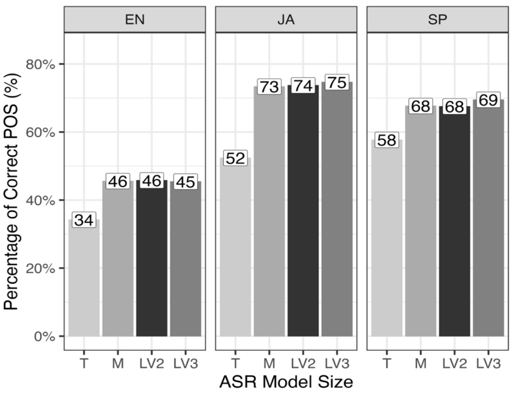


Figure 1: *Percentage of cases (and counts) of wrong words but with correct POS across all languages.*

The third layer of analysis distinguishes between LEX_er and GRAM_er, revealing patterns not captured by the other two layers (WER and POS_er). Figure 2 below presents the error rates broken down by language, model size, and word class (GRAMMATICAL or LEXICAL). The horizontal dotted line for each language represents the overall POS_er as reference.

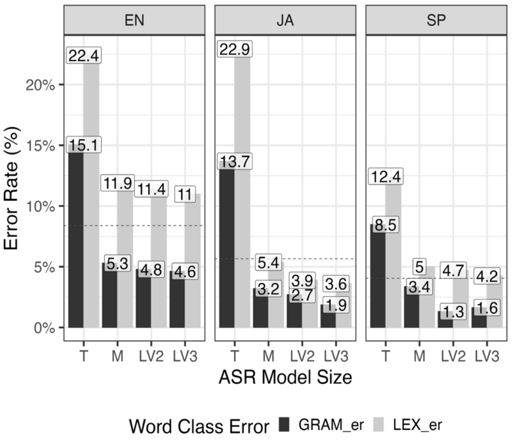


Figure 2: *Error rates across all languages and model sizes split by Word Class errors.*

Among the languages examined, Spanish consistently shows the lowest overall error rates, while English presents the highest. In the **LV3** model analysis, for LEX_er, Japanese records slightly lower rates than Spanish, while English exhibits the highest error rates (English = 11%; Japanese = 3.6%; Spanish = 4.2%). This variation can be explained linguistically by the fact that LEXICAL categories in Japanese and Spanish have higher inflections than in English, and these inflections are presented as affixes in both languages, helping the ASR system to understand the patterns of occurrence, useful to identify and predict the word form and its function in the language. This indicates that correct inflectional words significantly enhance predicting LEXICAL words. Although this finding is in contrast with [21], our results show that higher inflections are related to higher accuracy. Future research will help analyzing these results differences.

Further examination explored the extent to which predictable linguistic patterns helped in correctly identifying words. For this, we chose PROPER NOUNS (PROPN), as a subclass of the LEXICAL words. Our results show that Japanese is the language with least error rates, and English with the most errors (English = 39.6%; Japanese = 5.7%; Spanish = 18.6%). This is attributed to the use of case markers for PROPER NOUNS in Japanese, feature that is absent in English and Spanish, facilitating more accurate identification and prediction of this word class. The analysis revealed that the top six occurring words after PROPN were the case markers さん (3.4% – honorific particle), の (3.6% – possessive), に (2.4% – place), と (1.8% – joining nouns), は (1.8% – topic marking particle), and が (1.5% – grammatical subject), all accounting for approximately 15% of all words after PROPN in the Japanese dataset.

GRAM_er results show that Spanish had the lowest error rates compared to Japanese (slight difference of 0.3%) and, more prominently, to English (English = 4.6%; Japanese = 1.9%; Spanish = 1.6%). An in-depth analysis highlighted that the primary errors in English were associated with subordinating conjunctions (e.g., *if*, *that*, *while*) whereas the coordinating conjunctions were the ones driving more errors in Japanese (e.g., と *and*; も *also*) and Spanish (e.g., y *and*; o *or*). This indicates that a combination of grammatical assessment and linguistic function helps in a deeper understanding of how languages use specific words and the impact it has on the ASR accuracy. This approach is not necessarily language-dependent, but rather relies more on the typological function a word class has across multiple languages.

5. Discussion

This study discusses two interconnected error metrics to assess the accuracy of Whisper, using English, Japanese, and Spanish as test languages. Japanese and Spanish, being more inflectional than English, provided a valuable context for analysis, particularly in GRAMMATICAL words. Results reveal that relying solely on WER obscures important observations about ASR performance. In cases where WER percentages appear similar, as seen in the **T** model results, a closer examination at the GRAMMATICAL vs LEXICAL level unveils distinct accuracies. Conversely, even with differing WERs, such as in the **Large** models where Japanese and Spanish outperform English, a nuanced analysis exposes the ASR system's consistent performance on LEXICAL words but divergence in handling GRAMMATICAL words, with English being more prone to errors with the coordinating conjunctions. These metrics strike a balance between WER and individual POS errors. While reporting each POS category individually could complicate cross-language system performance comparisons, the analyzed metrics offer a middle ground. They enable users to identify the strengths and weaknesses of ASR at crucial linguistic levels, providing clarity on areas requiring attention when refining outputs. This approach enhances the interpretability and practical utility of ASR performance assessments.

6. Conclusions

The automatic processing and annotation of natural speech are complex tasks influenced by both the systems themselves and, most importantly, by the inherent characteristics of languages. Current systems have made significant progress in addressing these complexities. One notable advancement is the ability to perform automatic grammatical error comparisons across languages with different typological classifications. This advancement requires a cautious approach to understanding intrinsic language differences and variations based on the ASR system or the data used for training.

7. Acknowledgements

We would like to thank Hayden Ooi and Chloe Dean for their valuable feedback on an earlier version of this paper. We also wish to thank the anonymous reviewers for their helpful comments on the shape and content of the submission.

8. References

[1] O'Shaughnessy, D., "Trends and developments in automatic speech recognition research", Computer Speech & Language, vol. 83, 2023.

[2] Reitmaier, T. et al, "Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers", Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22), New York: USA, 299, 1-17, 2022.

[3] Whetten, R. and Kennigton, C., "Evaluating and Improving Automatic Speech Recognition using Severity", The 22nd Workshop on Biomedical NLP and BioNLP Shared Tasks, 79-91, 2023.

[4] Kumalija, E. and Nakamoto, Y., "Performance evaluation of automatic speech recognition systems on integrated noise-network distorted speech", Frontiers in Signal Processing, vol. 2, 1-10, 2022.

[5] He, X., Deng, L., and Acero, A., "Why word error rate is not a good metric for speech recognizer training for the speech translation task?", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5632-5635, 2011.

[6] Kafle, S. and Huenerfauth, M., "Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing", Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility, 165-174, 2017.

[7] Garofolo, J. S., Voorhees, E. M., Auzanne, C. G. P., Stanford, V. M., and Lund, B. A., "1998 TREC-7 spoken document retrieval track overview and results", Proceedings of the 7th Text REtrieval Conference. NIST, 79-89, 1998.

[8] Roux, T. B., Rouvier, M., Wottawa, J., and Dufour, R., "Qualitative evaluation of language model rescoring in automatic speech recognition", Proceedings INTERSPEECH 2022 - 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 3968-3972, 2022.

[9] Morris, A. C., Maier, V., and Green, P., "From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition", Proceedings INTERSPEECH 2004 - 8th Annual Conference of the International Speech Communication Association, Jeju Island, Korea, 2765-2768, 2004.

[10] Kheddar, H., Himeur, Y., Al-Maadeed, S., Amira, A., and Bensaali, F., "Deep transfer learning for automatic speech recognition: Towards better generalization", Knowledge-Based Systems, vol. 277, 2023.

[11] Lee, S., Noh, H., Lee, K., and Geunbae Lee, G., "Grammatical error detection for corrective feedback provision in oral conversations", Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI'11), AAAI Press, 797-802, 2011.

[12] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I., "Robust speech recognition via large-scale weak supervision", Proceedings of the 40th International Conference on Machine Learning, 28492-28518, 2023.

[13] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G., "Common Voice: A Massively-Multilingual Speech Corpus", Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 421-4215, 2020.

[14] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "Bert: Pretraining of deep bidirectional transformers for language understanding", North American Chapter of the Association for Computational Linguistics, 2019.

[15] Reimers, N. and Gurevych, I., "Sentence-bert: Sentence embeddings using siamese bert-networks", Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 3973-3983, 2019.

[16] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y., "Bertscore: Evaluating text generation with bert", 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.

[17] Popović, M. and Ney, H., "Word Error Rates: Decomposition over POS classes and Applications for Error Analysis", Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, 48-55, 2007.

[18] Hemchua, S. and Schmitt, N., "An Analysis of Lexical Errors in The English Compositions of Thai Learners", Prospect: An Australian Journal of TESOL, vol. 21(3), 3-25, 2006.

[19] McCowan, I. A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., and Bourlard, H., "On the use of information retrieval measures for speech recognition evaluation", Technical Report. IDIAP, 2004.

[20] Ethnologue: languages of the world, Dallas:Texas, SIL International, 1999.

[21] Berg, K., Hartmann, S., and Claeser, D., "Are some morphological units more prone to spelling variation than others? A case study using spontaneous handwritten data", Morphology, 2023.

[22] Lee, J. and Seneff, S., "Correcting Misuse of Verb Forms", Proceedings of ACL-08: HLT, Columbus, Ohio, USA, June 2008, Association for Computational Linguist, 174-182.

[23] Hisamitsu, T. and Nitta, Y., "An Efficient Treatment of Japanese Verb Inflection for Morphological Analysis", International Conference on Computational Linguistics, 1994.

[24] Centeno, J. G. and Obler, L. K., "Agrammatic verb errors in Spanish speakers and their normal discourse correlates", Journal of Neurolinguistics, vol. 14, 349-363, 2001.

[25] Park, C., et al., "Fast word error rate estimation using self-supervised representations for speech and text", arXiv preprint arXiv:2310.08225, 2023.

[26] R Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna:Austria, URL https://www.R-project.org/, 2021.

[27] Wijffels, J., Straka, M., and Straková, J., "Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the Udpipe Nlp Toolkit", https://CRAN.R-project.org/package=udpipe, 2021.

Application of ASR to a Sociolinguistic Corpus of Australian English

Maya Weiss[1], *Ksenia Gnevsheva*[1], *Catherine Travis*[1], *Gerard Docherty*[2]

[1] Australian National University, [2] Griffith University

maya.weiss@anu.edu.au, ksenia.gnevsheva@anu.edu.au, catherine.travis@anu.edu.au, gerry.docherty@griffith.edu.au

Abstract

This study applies Automatic Speech Recognition (ASR) to a sociolinguistic corpus of Australian English. We compare a human transcription of excerpts from 20 urban and regional speakers with a transcription generated by Microsoft's Azure AI Speech. The Word Error Rate is comparable to previous studies, and is not impacted by the sociolinguistic variables of speaker region and gender, nor the phonetic variable of vowel formants. Despite the overall low rate of transcription errors, our findings suggest that the quality of certain vowel categories that are particularly characteristic of Australian English can impact on the accuracy of the ASR-generated transcription.

Index Terms: Automatic Speech Recognition, corpus-building, transcription, sociolinguistics, Australian English

1. Introduction

Corpus building is central to much linguistic work, not least for sociolinguistic studies which are routinely centred around large samples of spontaneous speech. The creation of such corpora, however, is a notoriously slow process, due to the time required for both collecting speech data and producing usable transcriptions. Recent advances in technology have revolutionised transcription with the incorporation of ASR into the linguistics toolkit facilitating the processing of hours of speech in just minutes. There has been some work supporting the usability of ASR for sociophonetic analysis of American English [1], but there has been much less done with Australian English.

While the proprietary nature of ASR systems means that the nature of the material on which they are trained is opaque, a recurrent question in recent years has been how well ASR systems deal with cross-accent variation [2-5]. In the case of Australian English, with its well-documented differences from other varieties [6, 7] and its relatively small population of speakers, it is not a given that ASR-generated transcription would work as effectively as it does for other varieties (e.g. from the USA) which are likely to be more highly represented in training materials. For example, vowels with a realisation that is particularly distinctive for Australian English speakers may be prone to transcription errors, and the effectiveness of automatic transcription may be impacted by factors such as region and gender which are known to correlate with accent variability in Australian English varieties.

In this paper, we test Microsoft's Azure AI Speech [8] on a substantial sample of spontaneously spoken Australian English. We use the measure of Word Error Rate to assess accuracy, and we consider this for the data overall, across genders, and across speakers from an urban vs. regional setting. We further consider the impact of the acoustic signal, to compare accuracy across different vowel categories, including those that are more characteristically Australian and those that are not.

2. Automatic Speech Recognition

Current ASR models use combinations of deep learning systems. Typical components include signal processing and feature extraction, acoustic and language models, and a hypothesis search [9]. Together, these convert a speech signal into a frequency-domain representation; extract and match salient features with acoustic and phonetic knowledge; make grammatical and lexical-semantic predictions; and estimate the probabilities of hypothesised word sequences, outputting a word sequence with the highest probability. One salient feature of ASR systems is that they do not rely on segment-by-segment analysis of the acoustic phonetic parameters conventionally used in phonetic analysis, such as formant estimates.

A widely used method for testing ASR accuracy is *Word Error Rate* (WER), which is the sum of three types of errors (substitution, deletion, and insertion) divided by the total number of words in the corrected transcript [3]. Accuracy is impacted by the ASR model used, and the data fed into the system, including audio quality [10]. For example, OpenAI's Whisper achieved an average 12.8% WER across multiple English datasets, with the lowest WER (2.7%) being achieved on read speech [11]. Microsoft's 2017 system achieved an overall 5.1% WER with data from telephone conversations [12]. This is indicative of the fact that ASR models are better at transcribing what can be considered 'clean' data, reinforced by Meta (then Facebook) AI's wav2vec 2.0 model achieving a 1.8% WER on test-clean speech in 2020 [13].

ASR performance differs across varieties of a language. For example, Whisper ASR performed better with American and Canadian English over Australian and British English [14], and YouTube's automatically generated captioning performed better with American than Scottish English [4]. Within American English, several ASR systems were found to be more accurate for the speech of white Americans than of African Americans [3], and of white Americans than Native Americans, African Americans and ChicanX [5]. Similarly, ASR performed better with Standard Dutch than Southern Dutch Dialects [15]. Results for gender vary, with better performance observed for men [4], for women [14], and for neither [16]. Sociophonetic variation seems to play a role in accuracy, and one study of ASR for speakers of the Pacific North-West (USA) pointed to sociolinguistic features of that variety accounting for 20% of word errors [5].

One of the factors impacting ASR performance is the imbalanced nature of training data sets, which are believed to over-represent men and university-educated speakers [4] and to under-represent minoritised speakers [5]. It has been demonstrated that accuracy improves with dialect-specific training sets [17], and with larger training sets [11]. The fact that "stratified sampling of speakers has historically not been the priority during corpus construction for computational applications" [4] is thus potentially problematic for ASR systems, including in their use for sociolinguistic analysis.

3. Method

3.1. Data

The data in this study come from two corpora of Australian English, one recorded in the large metropolis of Sydney (pop. 5.2 million) [18], and the other in a regional district around the town of Braidwood, New South Wales (pop. 1720; 300km from Sydney and 100km from Canberra) [19]. The recordings are from casual interviews – sociolinguistic interviews for the urban data and oral history interviews for the regional data. Participants include five men and five women from each region, who are generally comparable in other social characteristics: they are of Anglo-Celtic background, aged between 35 and 65, and have a range of occupations (from lawyers to bar workers).

The analysis is based on 20,000 words, 1,000 words from each speaker (as counted in the manually corrected transcripts). We endeavoured to extract continuous, uninterrupted stretches of speech as much as possible, though due to the interactive nature of the recordings, it was necessary to extract multiple strings to reach 1,000 words. The sociolinguistic interviews are more conversational than the oral histories, thus for the former, the 1,000 words came from an average of 7.2 strings, and for the latter, from an average of 4 strings.

3.2. ASR Process

For this study, we used the ASR operated via Microsoft Azure AI Speech. The most recent detailed description of its architecture is from 2017 [12]. This does not specify the training data, although corpora such as Switchboard [20], LibriSpeech [21], and GigaSpeech Middle [22] are likely to be included [12, 23]. There is an option to set the language variety to Australian English (among others), implying that there is some Australia-specific information in the system.

Audio files were transcribed by Azure AI Speech via a dedicated speech-to-text transcription platform developed at Griffith University using Microsoft technologies and run locally. The language was set to Australian English, and the number of speakers was specified (two for all files but one with three speakers, including one interviewer, whose speech is not analysed). The recordings were processed, outputting .txt and .json transcription files, as well as a numerical rating (on a 0-1 scale) of the confidence with which the ASR had reached the corresponding output for each individual word and each phrase.

3.3. Data Coding

The ASR output was manually corrected, and these corrected transcripts were then compared word-by-word with the original ASR output to identify errors. Errors were coded for type (substitution, deletion, insertion) and linguistic element affected (vowel, consonant, filler, lexical, or other; see Table 1). We excluded proper nouns that were specific to the recording sites (e.g. *Araluen, Mongarlowe*).

Example (1) provides an illustration of ASR output for a segment of text and (2) the corresponding corrected transcription. Substitutions are underlined (e.g. *points, pawns* and *palms* for *pines*) and deletions are bolded (e.g. *all*, repetition of *the*, *um*). The overall WER in this sample is 17% (9 errors divided by 52 words in the corrected transcription). A third type of error not illustrated here is insertion, where the ASR output contains an addition to what was produced. Multiple errors are sometimes coded for the one word (e.g. ASR *go on* for *gone* has an insertion and a substitution; ASR *annoying* for *and I went* has two deletions and a substitution).

(1) ASR output
So we had points result ***Ø*** *here that* ***Ø*** *the house, we had two hectares all up for pawns. So right here at the house, there's some right up against the house and then the hill* ***Ø*** *just at the back of the paddock there, that was full of palm trees as well.*

(2) Corrected transcription
So we had pine trees ***all*** *here at* ***the*** *the house, we had two hectares all up of pines. So right here at the house, there's some right up against the house and then the hill* ***um*** *just at the back of the paddock there, that was full of pine trees as well.*

To assess the effect of vowel quality on transcription accuracy, the corrected transcripts were force-aligned to the corresponding speech signal (with LaBB-CAT [24]). F1 and F2 values for the primary/stressed vowel in the target word were estimated (with Praat [25]), at 20% into a diphthong interval and 50% for monophthongs. The F1/F2 estimates were subsequently mapped to vowels transcribed correctly and incorrectly (coded as 'vowel errors'). A total of 22,580 vowel tokens were processed by LaBB-CAT, including 360 vowel error tokens, of which 299 are analysed below (17% being set aside, due to processing issues in Praat or LaBB-CAT).

4. Results

4.1. Error Types and Word Error Rate

As a first measure of performance, we consider the Word Error Rate, and types of errors. Table 1 give the numbers of words impacted for both error types and affected linguistic elements, along with examples of each. SUB+ represents instances where multiple errors are involved (all of which involve a substitution and one or two other errors). As can be seen, substitutions and deletions are equally frequent, while insertions are relatively rare. The most common type of deletion is fillers such as *um, uh*. Though the ASR system used here did not categorically delete all fillers, many ASR models intentionally delete these [26], and thus it may be misleading to consider them errors. If we set fillers aside, then substitutions are overwhelmingly the most common type of error, of which vowel substitutions are particularly frequent.

The total number of errors is 1,420, giving an overall WER of 7%. Individual speakers exhibit a wide range, from 1.6% to 14.7%, with a per speaker mean of 7.2%. An Interquartile Range (IQR) calculation reveals no outliers. Excluding fillers, the overall WER drops to 5.3%, and the range for individual speakers drops to 0.6% to 9.9%, with a mean of 5.3% (again, no outliers, according to an IQR calculation).

Table 1: *Error types and affected linguistic elements.*

Error Type	Linguistic Element	Target word	ASR output	N
SUB (N=520)	Vowel	*known*	*nine*	269
	Consonant	*embers*	*members*	216
	Other	*threw*	*through*	35
DEL (N=570)	Filler	*um*	*Ø*	372
	Lexical	*then*	*Ø*	198
INS (N=26)	Lexical	*Ø*	*it's*	8
	Other	*rem~ ('remember')*	*mean*	18
SUB+ (N=304)	Vowel	*a firey*	*Afari*	91
	Other	*Australians*	*this train*	213

4.2. Word Confidence Scores

The confidence with which the ASR has reached a corresponding output is potentially a useful tool for researchers – if the incorrectly transcribed words are given a lower confidence rating than the correctly transcribed words, these could help to locate transcription errors. Table 2 provides these confidence scores and shows that correctly transcribed words do have a higher average score than incorrect words (0.808 vs. 0.595). The range, however, is similar across the two: for correct words, from 0.004 to 0.998, and for incorrect words from 0.012 to 0.984. Thus, in some cases, ASR outputs (whether correct or incorrect) are returned with low levels of confidence, and some instances of incorrect outputs are returned with high levels of confidence.

Table 2. *Confidence scores of correctly and incorrectly transcribed words.*

Error	Average Conf.	StdDev	Min. Conf.	Max. Conf.
Correct	0.808	0.179	0.004	0.998
Incorrect	0.595	0.245	0.012	0.984
Grand Total	0.798	0.189	0.004	0.998

Note: Confidence values come from uncorrected transcripts (total N words = 20,371; N correct = 19,718, N incorrect = 653.

4.3. Social Effects: Region and Gender

We found no significant difference in WERs for the two social factors considered. The WER for the urban speakers is slightly higher than for the regional speakers (6.1% vs. 4.5%; urban Mean = 6.1%, SD = 0.030; regional Mean = 4.5%, SD = 0.022; fillers excluded), but this difference is not significant (based on a two-sample t-test performed in Python, p = 0.200). The regional IQR is 3.8%, and the urban IQR is 4.9%, signifying a somewhat higher level of variation within urban speakers.

For gender, though men have a higher overall WER than women (6.7 vs. 4.4%; Men Mean = 6.2%, SD = 0.022; Women Mean = 4.4%, SD = 0.029; fillers excluded), this difference is not significant (p = 0.122). The female IQR is 4.5%, and the male IQR is 4.3%, thus little difference between the groups.

4.4. Acoustic Effects

We now turn to consider what impact vowel quality has on the ASR performance, for which we focus on the 10 most frequent vowel categories to occur in the data.

We first compare the acoustic properties of the correctly vs. incorrectly transcribed vowels, shown in Figure 1 with diphthongs in the top panel and monophthongs in the bottom panel (8 points with F1>1,100Hz were removed to aid with figure readability). The ellipses represent the correctly transcribed instances, restricted to the central 1SD, and the individual points of the same colour the incorrectly transcribed instances of the same vowel categories. While the incorrect vowels appear to be widely distributed, the majority fall within the ellipse of their respective category. For example, for PRICE 60% of the incorrectly transcribed tokens fall inside the ellipse (comparable to the 67% of correctly transcribed tokens which fall within 1 SD). Furthermore, acoustic characteristics of vowels that were incorrectly transcribed are not necessarily predictive of what they would be mistranscribed as. For example, the incorrectly transcribed instance of a PRICE vowel with the lowest F1 value (the top-most purple point in the diphthong chart, in the word *I*) was incorrectly transcribed as TRAP (*and*), whereas one would expect vowels with relatively high F1 and F2 values to be transcribed as TRAP. The two incorrectly transcribed instances of the FACE vowel with the highest F1 and F2 values (the bottom- and left-most red points, in the words *they* and *mates*) were incorrectly transcribed as GOAT and STRUT (*though* and *mums* respectively).

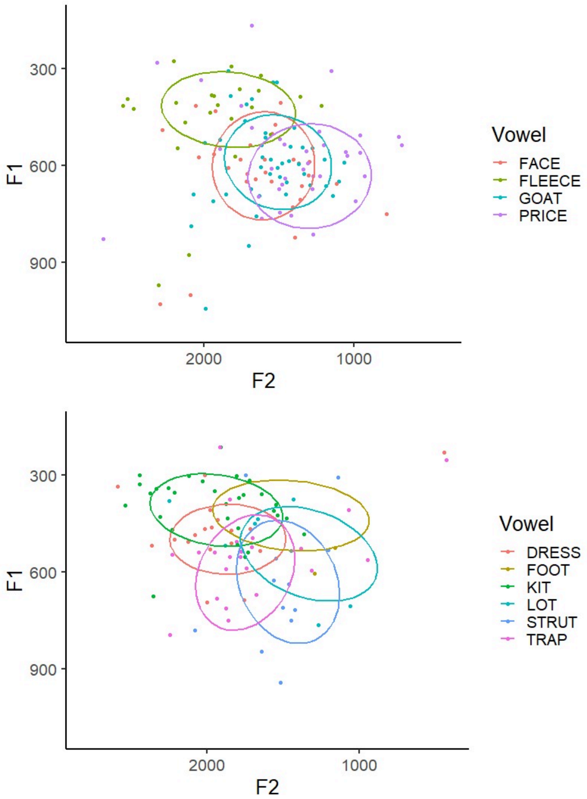


Figure 1. *Diphthongs and monophthongs correctly (ellipses) and incorrectly (points) transcribed by ASR.*

The fact that relative location within F1/F2 space does not emerge as a compelling predictor of the vowel substitution errors is of course not unexpected given that in ASR multiple channels of information cutting across different temporal domains are processed in parallel, well beyond the segment-based F1/F2 spectral specifications typically applied within acoustic phonetic analysis [2]. Nevertheless, it is interesting to note that the diphthongs tend to have a lower rate of accuracy than the monophthongs, as seen in Table 3, which gives the proportion of instances that are incorrect for these 10 vowels. The highest proportion of errors is with the diphthongs (in particular, GOAT, e.g. *so* as *sorry, inside, a*), while LOT has the lowest proportion of errors (e.g. *was* as *is*).

Table 3. *Proportion of errors per vowel category.*

Vowel category	% incorrect	N
GOAT	3.5%	1238
FACE	2.7%	1131
PRICE	2.6%	1405
FLEECE	1.5%	1681
TRAP	1.4%	1730
DRESS	1.4%	1358
STRUT	0.9%	1497
KIT	0.8%	4024
FOOT	0.8%	1444
LOT	0.6%	1424

A final point is that there is a cluster of vowel substitution transcription errors (77 in total; 21% of the vowel substitution errors) that *prima facie* do point to the ASR analysis being influenced by specific distinctive segmental characteristics of the Australian English vowel system. Table 4 presents some examples of where the ASR output has conceivably been triggered by the typical realisation of a different vowel category by a speaker within our sample. This cluster was not distributed evenly across vowel categories with the majority of instances being found in lexical sets where there is not a good match in phonetic realisation between Australian English and those other varieties that likely constitute the larger component of the ASR training materials. Likewise, this cluster of apparently Australian English vowel-triggered errors was made up of tokens from some but not all speakers, reflecting the fact that within our speaker sample there will inevitably be significant variability in the way in which some of the same vowel categories are realised. Though it is not known the extent to which segmental characteristics are influential in the decisions made by ASR transcription algorithms, examples such as these do suggest that they are relevant to differences in ASR performance across varieties of the same language, albeit only at the margins.

Table 4. *Examples of ASR errors with vowels characteristic of Australian English.*

Vowel category	Target word	ASR Output
FACE > PRICE	*trades*	*tried*
	wave	*why*
FLEECE > FACE	*seen*	*saying*
	green	*grain*
	meal	*mail*
GOAT > STRUT	*ropes*	*rubs*
PRICE > CHOICE	*aisles*	*soils*
STRUT > TRAP	*tugs*	*tags*
(pre-nasal) TRAP > FACE	*plans*	*planes*
(non-rhotic) BATH > TRAP	*bark*	*back*

5. Discussion

This study has explored ASR transcriptions produced by Microsoft's Azure AI Speech for Australian English spontaneous speech, considering WERs, error types, and confidence scores, and testing whether accuracy varies according to social (gender and region) and phonetic factors (formant values).

First, overall WER was 7% with fillers included, and 5.3% with fillers removed. This WER is very close to the 5.1% reported by Microsoft for conversational American English [12]. Though this might suggest that the ASR does not perform worse with Australian English, this study is now seven years old and, given rapid advances in technology, the comparison may not be a valid one. We considered the confidence scores at the word level as a potential tool for researchers to pinpoint areas for correction, but these proved not to be a reliable measure of correct ASR transcription. The wide range of confidence values not only in incorrect words but also in correct words indicates that the ASR does not always recognise when it is making an error.

Looking more specifically at the kinds of errors that occurred, we found that one of the most frequent errors was filler deletions. Although ASR models often purposefully exclude fillers [26], this was not the case here where, despite filler deletion being the most frequent error, *um* and *uh* were nevertheless transcribed in the ASR output multiple times, particularly when said in relative isolation, separated by pauses. It may be that ASR models are trained on clean speech that contains few fillers [27], and thus they cannot reliably be detected, particularly if not said in isolation. Whether fillers are regularly deleted or not, however, is more reliant on the specific ASR model being used. While filler deletions are less problematic for sociophonetic analysis, the other most frequent error is problematic, namely vowel substitutions, which we hypothesised may be related to a lack of adequate representation of Australian varieties of English in datasets used to train ASR models.

Our sociolinguistic comparisons revealed no significant differences, although women had a lower WER than men, and regional speakers had a numerically lower WER than urban speakers. The lack of a significant effect for gender may not be surprising, given the varied results from previous studies [4, 14, 16]. The lack of a significant effect for region, however, is contrary to what we might predict, if the regional speakers accord with the stereotype of using more characteristically Australian English vowels [28] or follow a general trend of lying behind language changes taking place in urban centres [29]. More work is required across varieties of Australian English, including regional varieties to help shed light on this.

Finally, we explored whether the phonetic qualities of vowel segments influenced the accuracy of the ASR output. We found that this was not the case, with segments that were incorrectly transcribed showing a similar spread of vowel formants to those that were correctly transcribed, likely attributable to the fact that ASR does not rely on formant frequencies to identify speech. However, there was an indication that Australian English vowel realisations may have been problematic for the ASR, seen in poorer performance with diphthongs and other vowel categories that are particularly distinctive for Australian English.

6. Conclusion

While the scope of this study is constrained by a relatively small dataset (20,000 words from 20 speakers), our findings provide clear evidence that using an ASR approach to generating automatic orthographic transcription can work effectively with unscripted speech produced by a diverse sample of Australian speakers of English. Furthermore, the accuracy of the transcription does not appear to be influenced by factors that are known to impact on the characteristics of the speakers' spoken performance such as gender or whether they are based in an urban or regional location. The fact that the ASR model is deploying information of a form and complexity that is some way removed from conventional acoustic phonetic parameters explains why vowel location in F1/F2 space is not a strong predictor of whether a transcription instance is correct or incorrect. There is however a suggestion that some incorrect transcriptions are arising as a result of the ASR model mis-interpreting vowel qualities that are known to be particularly distinctive in the speech of (some) Australian speakers of English. Testing a wider range of speakers and accents will allow us to better understand the factors that are most closely associated with lower levels transcription accuracy, thus facilitating more informed use of ASR for sociolinguistic work.

7. Acknowledgements

We thank Clare Young and the Braidwood Museum for making the interviews from the "Heart of the Storm" podcast series available for this research; Gan Qiao for assistance with data management for the ASR process; and an anonymous reviewer for comments on an earlier version of this paper. This work was supported by ARC DP230100464, and was initiated as a Summer Scholar project conducted by the first author, with support from the Commonwealth Government of Australia.

8. References

[1] Coto-Solano, R., Stanford, J. N., and Reddy, S. K., "Advances in completely automated vowel analysis for sociophonetics: Using end-to-end speech recognition systems with DARLA," Frontiers in Artificial Intelligence, 4, 2021-September-24, 2021. https://doi.org/10.3389/frai.2021.662097.

[2] Coto-Solano, R., "Computational sociophonetics using Automatic Speech Recognition," Language and Linguistics Compass, 16(9):e12474, 2022. https://doi.org/10.1111/lnc3.12474.

[3] Koenecke, A. et al., "Racial disparities in automated speech recognition," Proceedings of the National Academy of Sciences, 117(14):7684-7689, 2020. https://doi.org/10.1073/pnas.191576811.

[4] Tatman, R., "Gender and dialect bias in YouTube's automatic captions," Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Valencia, Spain, 53-59, 2017. DOI: 10.18653/v1/W17-1606.

[5] Wassink, A. B., Gansen, C., and Bartholomew, I., "Uneven success: Automatic Speech Recognition and ethnicity-related dialects," Speech Communication, 140:50-70, 2022/05/01/, 2022. https://doi.org/10.1016/j.specom.2022.03.009.

[6] Cox, F., and Fletcher, J., Australian English pronunciation and transcription, 2nd ed., Cambridge: Cambridge University Press, 2017

[7] Purser, B., Grama, J., and Travis, C. E., "Australian English over time: Using sociolinguistic analysis to inform dialect coaching," Voice and Speech Review, 14(3):269-291, 2020. https://doi.org/10.1080/23268263.2020.1750791.

[8] Microsoft. "Azure AI Speech," Accessed: 23 September 2024. https://azure.microsoft.com/en-us/products/ai-services/ai-speech/.

[9] Yu, D., and Deng, L., Automatic Speech Recognition: A Deep Learning Approach, London: Springer, 2014. https://doi.org/10.1007/978-1-4471-5779-3.

[10] Loakes, D., "Does Automatic Speech Recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes?," Frontiers in Communication, 7, 2022-June-14, 2022. https://doi.org/10.3389/fcomm.2022.803452.

[11] Radford, A. et al., "Robust speech recognition via large-scale weak supervision," International Conference on Machine Learning, 28492-28518, 2023. https://doi.org/10.48550/arXiv.2212.04356.

[12] Xiong, W. et al., "The Microsoft 2017 conversational speech recognition system," Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5934-5938, 2018. https://doi.org/10.1109/ICASSP.2018.8461870.

[13] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M., "wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in Neural Information Processing Systems, 33:12449-12460, 2020. https://doi.org/10.48550/arXiv.2006.11477.

[14] Graham, C., and Roll, N., "Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits," JASA Express Letters, 4(2), 2024. https://doi.org/10.1121/10.0024876.

[15] Ghyselen, A.-S. et al., "Clearing the transcription hurdle in dialect corpus building: The corpus of Southern Dutch dialects as case study," Frontiers in Artificial Intelligence, 3, 2020-April-15, 2020. https://doi.org/10.3389/frai.2020.00010.

[16] Tatman, R., and Kasten, C., "Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube Automatic Captions," Interspeech, 934-938, 2017. https://10.21437/Interspeech.2017-1746.

[17] Dorn, R., "Dialect-specific models for Automatic Speech Recognition of African American Vernacular English," Proceedings of the Student Research Workshop Associated with RANLP 2019, Varna, Bulgaria, 16-20, 2019. DOI: 10.26615/issn.2603-2821.2019_003.

[18] Travis, C. E. et al., "*Sydney Speaks Corpus*," 2023. https://dx.doi.org/10.25911/m03c-yz22.

[19] Travis, C. E., Gnevsheva, K., and Docherty, G., "Voices of Regional Australia Corpus," In Progress

[20] Godfrey, J. J., Holliman, E. C., and McDaniel, J., "SWITCHBOARD: Telephone speech corpus for research and development," Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 517-520, 1992. DOI: 10.1109/ICASSP.1992.225858.

[21] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S., "Librispeech: An ASR corpus based on public domain audio books," Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5206-5210, 2015. DOI: 10.1109/ICASSP.2015.7178964.

[22] Chen, G. et al., "Gigaspeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," Interspeech, 3670-3674, 2021. http://dx.doi.org/10.21437/Interspeech.2021-1965.

[23] Gong, X. et al., "Advanced long-content speech recognition with factorized neural transducer," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32:1803-1815, 2024. https://doi.org/10.1109/TASLP.2024.3350893.

[24] Fromont, R., and Hay, J., "LaBB-CAT: An annotation store," Proceedings of the Australasian Language Technology Workshop:113-117, 2012. https://aclanthology.org/U12-1015.

[25] Boersma, F., J., and Weenink, D., Praat: Doing phonetics by computer [Computer program] Version 6.4.13: retrieved 10 June 2024 from http://www.praat.org/, 2024

[26] Lease, M., and Johnson, M., "Early deletion of fillers in processing conversational speech," Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, New York City, USA, 73–76, 2006.

[27] Zhu, G., Caceres, J.-P., and Salamon, J., "Filler word detection and classification: A dataset and benchmark," Interspeech, 3769-3773, 2022. https://doi.org/10.48550/arXiv.2203.15135.

[28] Bradley, D., "Regional characteristics of Australian English: Phonology," Varieties of English: The Pacific and Australasia, Burridge, K. and Kortmann, B., eds., 111-123, Berlin/New York: Mouton de Gruyter, 2008. https://doi.org/10.1515/9783110208412.1.111.

[29] Britain, D., "Space and spatial diffusion," The Handbook of Language Variation and Change, Chambers, J. K. et al., eds., 603-637, Malden, MA: Blackwell, 2004. https://doi.org/10.1002/9780470756591.ch24.

Does the Inclusion of Other Modalities Enhance the Performance of Speech Emotion Recognition Systems?

Junchen Liu[1], Jesin James[1], Karan Nathwani[2]

[1]The University of Auckland, [2]India Institute of Technology Jammu

jliu522@aucklanduni.ac.nz, jesin.james@auckland.ac.nz, karan.nathwani@iitjammu.ac.in

Abstract

The pursuit of natural human-computer interaction has driven the advancement of emotion recognition technology. Speech emotion recognition (SER) has gained widespread attention due to its high applicability. Recently, some researchers have been interested in developing multi-modal emotion recognition (MER) systems that integrate speech with text and video modalities to enhance robustness and accuracy. We analyse the performance of these systems using the IEMOCAP and RAVDESS datasets, highlighting the impact of different modality combinations on emotion recognition accuracy. This paper aims to guide future research in optimising MER by leveraging the complementary advantages of various modalities.

Index Terms: speech emotion recognition, multi-modal emotion recognition, impact of different modality combinations

1. Introduction

Emotion recognition (ER) technology could improve the feasibility of human-computer interaction in real-world applications; it enables computers and other intelligent devices to understand and analyse the emotional state of users, thereby providing personalised and humanised services. However, due to the diversity, complexity and subjectivity of human emotional expression, the implementation of ER technology faces numerous challenges and difficulties.

Speech emotion recognition (SER) systems have received significant attention due to their potential applications in various fields, such as mental health monitoring and customer service enhancement. In recent years, there has been increasing enthusiasm about incorporating text or video modalities into SER models to develop multi-modal emotion recognition (MER) systems. MER systems concurrently process information from multiple modalities, which may increase the stability of the system [1]. For example, when the speech modality is unable to effectively predict the emotional states due to the environmental noise, the intervention of video or text modality can provide valuable emotional information to the MER systems. In previous research on MER tasks, rarely studies have identified which modality contributes most to emotion recognition. However, it is crucial to investigate the impact of each modality on recognition accuracy. These explorations can enhance the understanding of the complementarity and contributions of each modality, thereby simplifying model design by removing unnecessary modalities, reducing system complexity and improving computational efficiency.

Adopting deep learning techniques, especially deep neural networks (DNN), has become a prevalent trend in ER research. DNN, characterised by deep network structures and a substantial number of parameters, can automatically learn complex representations and extract relevant features from input data, capturing subtle emotional cues that may be missed by hand-crafted features [2].

In summary, this paper's primary objective is to explore different modalities' contributions to the performance of MER systems and answer the following research questions:

1. What are the advantages and disadvantages of each modality (speech, video, text) in detecting various emotional states?
2. How does combining different modalities affect the overall performance of MER systems?

2. Literature Review

This section will elaborate on the datasets and the methods of the state-of-the-art ER systems.

2.1. Datasets

Researchers in the field of SER and MER often select English datasets such as the interactive emotional dyadic motion capture (IEMOCAP) [3] and ryerson audio-visual database of emotional speech and song (RAVDESS) [4] as their primary resources for model construction. These datasets are favoured due to the rich emotional content and multiple modalities, offering robust support for developing effective ER systems. The details of these datasets are shown in Table 1.

Table 1. *IEMOCAP and RAVDESS dataset's information. S, V, T, H, F represents speech, video, text, hand movement and facial expression respectively.*

Dataset	Modalities	Utterances	Emotions
IEMOCAP [3]	S, V, T, H, F	7529	9
RAVDESS [4]	S, V	1440	8

Specifically, the IEMOCAP dataset records the performance data of 10 actors (5 female, 5 male). It has 9 emotions, such as neutral (1708), happy (595), sad (1084), anger (1103), surprise (107), fear (40), disgust (2), frustrated (1849) and excitement (1041). In total, the database contains approximately twelve hours of data. Due to the fact that IEMOCAP is an unbalanced dataset, some researchers tend to use only four categories to construct balanced datasets; they combine happiness and excitement as the happy category, and the other three classes are anger, sadness and neutral.

RAVDESS dataset consists of 1440 utterances by 24 professional actors (12 female, 12 male), encompassing eight emotion types: calm, happy, sad, angry, fearful, surprised, disgusted, and neutral. RAVDESS is the balanced dataset, with 192 utterances for each emotion except for neutral, which has 96 utterances.

2.2. Speech Emotion Recognition

Acoustic information, such as tone, pitch, and frequency, contains rich emotional cues that can precisely express human emotional states. This information in speech forms the foundation for the application of SER systems in various fields, including intelligent customer service, personal assistants, and market research. However, the performance of SER systems is often affected by the environmental noise [5]. To address this issue, researchers are dedicated to developing SER systems that exhibit robust noise resistance, ensuring reliable and accurate emotion detection in diverse and noisy environments.

Convolutional neural networks (CNN) have gradually become the primary model for SER tasks. This is because CNN models can capture local and multi-level features with translation invariance in speech signals; these features will help improve the robustness and noise resistance of SER systems [6]. Most researchers choose to extract speech features from raw speech waveforms [7] and mel-spectrograms [8], or transfer the speech signal into the mel-frequency cepstral coefficients (MFCC) [9] as the input.

Additionally, some researchers illustrate that the transformer model based on DNN is suitable for processing long-time sequences of speech data. This is because the transformer model possesses a self-attention mechanism, enabling it to dynamically focus on the most relevant information at various positions in the input sequence, thereby improving the model's understanding of speech data and filtering out noise [10]. For example, Chen et al. [11] utilises Wav2Vec-2.0 to extract speech features from the raw speech waveform, and authors in [12] evaluate the effectiveness of different transformer models in SER, such as HuBERT, Wav2vec-2.0, and WavLM. In addition, the transformer-based model designed explicitly for log-mel spectrograms, as outlined in [13], has demonstrated high recognition accuracy in handling SER tasks.

2.3. Multi-modal Emotion Recognition

This section primarily introduces the feature extraction methods for text and video modalities in MER systems, as well as the techniques for feature fusion. The feature extraction method for the speech modality remains consistent with the approach described in the previous section and will not be reiterated here.

Text data has the characteristics of accessible collection, storage, and processing. Meanwhile, different languages and cultures have specific ways of expressing emotions, text-based ER systems can utilise specific emotional vocabulary and phrases in these languages to improve the accuracy of emotion recognition. However, it is worth noting that polysemy in words can reduce the effectiveness of text-based ER systems [14]. Therefore, developing advanced text feature extraction methods that capture contextual information is essential for better understanding the meaning of the text.

The GloVe word vector model has been proven effective in text feature extraction. Rajan et al. [15] proposed a method that combines GloVe with the bidirectional gated recurrent unit (BiGRU) for text feature extraction, aiming to capture contextual dependencies in the text and enhance the model's ability to process long sequences. Additionally, embeddings from language models V2 (ELMoV2) [16], built on bidirectional long short-term memory (BiLSTM), provide context-aware word embeddings, meaning that the representation of each word dynamically changes based on its context within the sentence. This allows for better handling of polysemy and synonyms, enhancing the model's understanding capabilities. Currently, using transformer models for text feature extraction has garnered widespread attention. Transformer models, such as bidirectional encoder representations from transformers (BERT) [17] and robustly optimized BERT pretraining approach (RoBERTa) [18], possess self-attention mechanisms that can capture global dependencies between words in a text. These models also provide efficient parallel processing capabilities and strong bidirectional context understanding. These advantages make transformer models excel in text feature extraction tasks.

Facial images extracted from videos are commonly used as inputs for video feature extraction. Facial expressions are a direct and natural way of expressing emotions, different emotional states, such as happiness, sadness and anger, are usually manifested through changes in facial muscles, which are easily captured and recognised. Additionally, some microscopic details that indicate changes in emotional states, such as eye movements or the corners of the mouth turning up or down, can only be captured through facial expression. Notably, most research on video-based ER is affected by the issue of facial occlusion [19].

Recently, methods for extracting video features mainly revolve around CNN models. For example, [20] and [21] utilises the ability of 3-dimensional CNN (3D CNN) models to simultaneously capture spatial and temporal features to process a series of consecutive frames, thereby understanding the changes in facial expressions and movements in videos. On the basis of the 3D CNN model, [22] adds long short-term memory (LSTM) to gain spatial and temporal information, thereby better understanding long-term emotional changes in the video. For 2-dimensional CNN (2D CNN) models, The residual block of the residual convolutional network (ResNet) enables it to simultaneously extract features from low-level to high-level, forming a hierarchical representation, [23] illustrates that this model can capture subtle emotional changes in facial images. In addition, [24] proposed a spatio-temporal convolutional neural framework to extract features from face images. Combining a deep spatial network and a deep temporal network makes it possible to simultaneously capture spatial and temporal features in images, thereby generating comprehensive feature representations.

Feature fusion plays a crucial role in MER systems. An effective feature fusion method can utilise the complementarity between different modalities to integrate the most relevant features and ignore redundant or noisy information, thereby improving the model's recognition accuracy and generalisation ability. In recent years, researchers have tended to use attention mechanism-based feature fusion methods to integrate features from different modalities. For example, [18] uses cross-modal attention to fuse speech and text features. Cross-modal attention allows the model to adaptively learn interactions and dependencies between modalities. By dynamically assigning weights to features based on their contributions, cross-modal attention effectively integrates information from multiple modalities. [15] compared the impact of using cross-modal attention and self-attention as feature fusion methods on the accuracy of the MER system that combines video, text, and speech.

Table 2 shows the accuracy of state-of-the-art MER systems, which contain bi-modal and tri-modal ER systems.

Table 2. *Accuracy for MER system in IEMOCAP and RAVDESS datasets. S, V, T represents the speech, video and text modality respectively.*

Model	Modalities	Dataset	emotions	UA
33 speech features + ELMo v2 [16]	S, T	IEMOCAP	4	0.745
MSRFG [18]	S, T	IEMOCAP	6	0.716
(2D CNN + RNN) + 3D CNN [20]	S, V	IEMOCAP	3	0.717
(2D CNN + GRU) + 3D CNN [21]	S, V	IEMOCAP	4	0.764
MCWSA-CMHA [23]	S, V, T	IEMOCAP	4	0.863
1D CNN + ResNet + GloVe [15]	S, V, T	IEMOCAP	7	0.642
RDesBert [17]	S, V, T	IEMOCAP	7	0.792
MRPN [25]	S, V	RAVDESS	8	0.914
DSN + DTN + 1DCNN [24]	S, V	RAVDESS	8	0.949
RDes [17]	S, V	RAVDESS	8	0.980

3. Methodology

This section primarily elaborates on the methodology for SER and speech-based MER systems. The model selection is based on one of the state-of-the-art MER systems, [17] proposes an approach to improving the accuracy and robustness of the MER system. By integrating advanced feature extraction techniques, including BERT, ResNet and DenseNet, with a novel feature fusion method, highlighting the importance of feature recalibration through squeeze-and-excitation (SE) blocks, which improves model generalisation ability and performance. This MER system addresses critical challenges in understanding and processing complex human emotions. The architecture of the ER systems is shown in Figure 1.

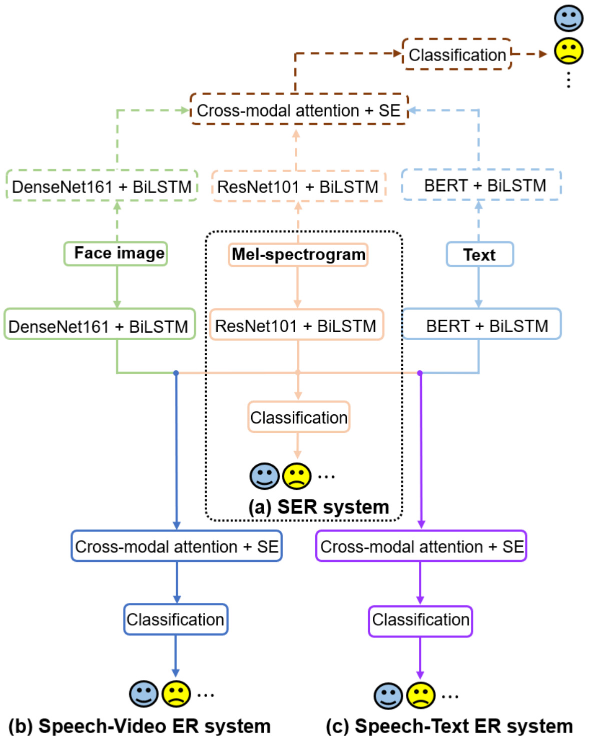


Figure 1: *Architecture of speech, speech-video, speech-text and speech-video-text ER systems.*

In the SER system, the original speech signals are converted into mel-spectrogram images. These images serve as the input for the combination of ResNet101 and BiLSTM. The aim of ResNet is to extract features from the mel-spectrogram, and the bi-directionality of BiLSTM allows the network to obtain past and future dependencies in the input sequence. ResNet101 consists of 5 convolutional regions with repeated convolutional calculations, totalling 101 convolutional layers. The architecture of SER is depicted in Figure 1 (a).

The input for the video modality in the speech-video ER system is facial images cropped from the video. The details are shown in Figure 1 (b). This bi-modal ER system combines DenseNet and BiLSTM to extract features from facial images. DenseNet, through its dense connection structure, achieves efficient feature transmission and reuse, providing robust feature extraction and generalisation capabilities, and the BiLSTM can further capture temporal information in the video. DenseNet161 has 161 convolutional layers.

The speech-text ER system employs BERT as the feature extractor for the text modality, which is shown in Figure 1 (c). The BERT architecture is composed of three main components: input embedding, transformer encoder, and output layer. Among these items, the most crucial part is the series of transformer encoder layers, designed to capture the contextual information between words within a sentence, thereby improving the system's ability to understand text content.

In terms of feature fusion methods, cross-modal attention and SE block are combined to fuse features from different modalities. The SE block can selectively enhance relevant information within each modality's features by assigning different weights to channels, thereby indirectly reducing the impact of redundant information on the model. Additionally, by weighing different features, the SE block enhances feature distinctiveness, making the system sensitive to subtle emotional changes and improving its generalisation capability.

The classification module integrates a statistical pooling layer, fully connected layers, and softmax activation. The statistical pooling layer can effectively capture the temporal dynamics of the input features by calculating statistical measures such as mean and standard deviation. Additionally, the statistical pooling layer integrates the fused features through these statistical measures, creating a comprehensive representation. This representation reflects information from all modalities, enabling the model to fully understand and utilise the input data. The combination of fully connected layers and the softmax activation function is a commonly used method in classification tasks. The fully connected layers extract and integrate features, while

Table 3. *Comparison between SER and MER systems in IEMOCAP and RAVDESS dataset. S, V, T represents the speech, video and text modality respectively.*

Model	Modalities	Dataset	emotions	WA	UA
ResNet101	S	IEMOCAP	7	0.468	0.479
BERT	T	IEMOCAP	7	0.535	0.551
DenseNet161	V	IEMOCAP	7	0.687	0.701
RBert	S, T	IEMOCAP	7	0.552	0.607
DesBert	V, T	IEMOCAP	7	0.734	0.769
RDes	S, V	IEMOCAP	7	0.742	0.788
RDesBert	S, V, T	IEMOCAP	7	0.756	0.792
ResNet101	S	RAVDESS	8	0.742	0.752
DenseNet161	V	RAVDESS	8	0.938	0.944
RDes	S, V	RAVDESS	8	0.963	0.980

the softmax activation function converts these features into class probabilities, thereby achieving classification prediction.

4. Results

We compared the weighted accuracy (WA) and unweighted accuracy (UA) of speech, text, video, speech-text, speech-video, video-text, and speech-video-text ER systems on the RAVDESS and 7-class unbalanced IEMOCAP datasets. The results are shown in the Table 3.

From the table, it could be observed that when using the 7-class IEMOCAP dataset as input, the UA of the speech-based and text-based ER system is only 47.3% and 55.1%, respectively. After incorporating the text modality into the SER system, the emotion recognition accuracy reaches 55.2% (WA) and 60.7% (UA). Despite this improvement, it has not surpassed the recognition accuracy of the video-based ER system. In contrast, the UA of the speech-video and text-video ER systems increased by 18.1% and 16.2% compared to the speech-text ER system.

For the speech-video-text ER system, the UA and WA increased by only 0.14% and 0.04% compared to the speech-video ER system. By analysing the experimental results shown in Table 3, it can be concluded that even though the accuracy of the text-based ER system is higher than that of the SER system, the contribution of the speech modality to the accuracy of the tri-modal ER system is more significant than that of the text modality. Therefore, the final contribution ranking is video, audio, and text. This result indicates that the video modality plays a more crucial role in providing emotional cues and enhancing overall recognition performance, while the addition of the text modality offers relatively minor improvements to the system's accuracy.

Since the RAVDESS dataset only provides speech and video modalities, it is impossible to determine the contribution of the text modality to recognition accuracy. However, the results presented in Table 3 indicate that the performance of the MER system, which combines video and speech, is superior to that of the SER system.

5. Discussion

Based on the results shown in Table 3, it can be observed that the recognition accuracy of the SER system for the 7-class IEMOCAP dataset is relatively low. Additionally, the accuracy gain achieved by incorporating the video modality into the SER system is much higher than that obtained by incorporating the text modality. The possible reason is that in an imbalanced dataset, the SER system may not thoroughly learn the features of the minority classes, leading the model to bias towards recognising the majority classes, thus affecting the overall recognition accuracy. Additionally, compared to video features, speech and text features are more dependent on temporal information, making them more susceptible to the effects of data imbalance.

Video modality provides rich and universal emotional expression information, which means that even if the amount of data for certain classes is relatively small, the system can still find cues to distinguish these classes from the limited features, making its performance relatively stable when facing imbalanced datasets. For example, compared to speech signals, certain emotions are more clearly and consistently expressed visually, such as smiles or frowns, which are consistent across different people and scenarios. Therefore, the model can learn these visual features easily, reducing its dependence on data balance.

The results of the MER system for the IEMOCAP dataset, as shown in Table 3, indicate that imbalanced datasets significantly impact the accuracy of emotion recognition. As the number of emotion categories increases, the imbalance becomes obvious, leading to a decrease in the recognition accuracy of the MER system. Notably, even with the 7-class imbalanced IEMOCAP dataset, the RDesBert system achieves an accuracy of over 79%. For the RAVDESS dataset, being a balanced dataset with frontal faces and no facial occlusions in the videos, MER systems typically achieve better results.

6. Conclusion

In conclusion, by leveraging the complementary strengths of different modalities, MER systems can achieve higher robustness and accuracy. Based on our research in the IEMOCAP dataset, incorporating video or text modalities into a SER system can improve the performance of emotion recognition. The video modality provides crucial emotional cues that significantly improve the system's ability to predict emotions precisely. While the text modality offers additional benefits, its contribution is comparatively less significant. Therefore, from the perspective of model complexity, it may be considered unnecessary to include the text modality if sufficient video and speech modality data are available. For future research, addressing challenges like data imbalance and improving feature fusion techniques are essential for the MER field.

7. References

[1] Zhao, S., Jia, G., Yang, J., Ding, G. and Keutzer, K., “Emotion Recognition from Multiple Modalities: Fundamentals and Methodologies”, IEEE Signal Processing Magazine, 38(6): 59–73, 2021.

[2] Sarma, M., Ghahremani, P., Povey, D., Goel, N., Sarma, K. and Dehak, N., “Emotion Identification from Raw Speech Signals Using DNNs”, in Interspeech, 3097–3101, 2018.

[3] Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S. and Narayanan, S., ”IEMOCAP: Interactive Emotional Dyadic Motion Capture Database”, Language resources and evaluation, 42: 335-359, 2008.

[4] Livingstone, S. and Russo, F., ”The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English”, Plos one, 13(5), 2018.

[5] Fahad, M., Ranjan, A., Yadav, J. and Deepak, A., “A Survey of Speech Emotion Recognition in Natural Environment”, Digital Signal Processing, 110, 2021.

[6] Trinh, V., Dao, T., Le, X. and Castelli, E., “Emotional Speech Recognition Using Deep Neural Networks”, Sensors, 22(4): 1414, 2022.

[7] Pandey, S., Shekhawat, H. and Prasanna, S., “Emotion Recognition from Raw Speech Using WaveNet”, in Region 10 Conference (TENCON), IEEE, 1292–1297, 2019.

[8] Zhao, J., Mao, X. and Chen, L., “Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks”, Biomedical Signal Processing and Control, 47: 312–323, 2019.

[9] Siadat, S., Voronkov, I. and Kharlamov, A., “Emotion Recognition from Persian Speech with 1D Convolution Neural Network”, in Fourth International Conference Neurotechnologies and Neurointerfaces (CNN), IEEE, 152–157, 2022.

[10] Alonazi, B., Nauman, M., Jahangir, R., Malik, M., Alkhammash, E. and Elshewey, A., “Transformer-based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion”, Applied Sciences, 12 (18): 9188, 2022.

[11] Chen, L. and Rudnicky, A., “Exploring Wav2Vec 2.0 Fine Tuning for Improved Speech Emotion Recognition”, in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023.

[12] Kakouros, S., Stafylakis, T., Mošner, L. and Burget, L., “Speech-based Emotion Recognition with Self-supervised Models Using Attentive Channel-wise Correlations and Label Smoothing”, in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023.

[13] Lu, C., Lian, H., Zheng, W., Zong, Y., Zhao, Y. and Li, S., “Learning Local to Global Feature Aggregation for Speech Emotion Recognition”, arXiv preprint arXiv:2306.01491, 2023.

[14] Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., Jia, W. and Yu, S., “A Survey on Deep Learning for Textual Emotion Analysis in Social Networks”, Digital Communications and Networks, 8(5): 745–762, 2022.

[15] Rajan, V., Brutti, A. and Cavallaro, A., “Is Cross-Attention Preferable to Self-Attention for Multi-Modal Emotion Recognition?” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 4693–4697, 2022.

[16] Singh, P., Srivastava, R., Rana, K. and Kumar, V., “A Multimodal Hierarchical Approach to Speech Emotion Recognition from Audio and Text”, Knowledge-Based Systems, 229, 2021.

[17] Liu, J., James, J. and Nathwani, K., “Improved Multi-modal Emotion Recognition Using Squeeze-and-excitation Block in Cross-modal Attention”, in Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2023.

[18] Wei, J., Hu, G., Tuan, L., Yang, X. and Zhu, W., “Multi-scale Receptive Field Graph Model for Emotion Recognition in Conversations”, in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023.

[19] Grahlow, M., Rupp, C. and Derntl, B., “The Impact of Face Masks on Emotion Recognition Performance and Perception of Threat”, PLoS One, 17(2), 2022.

[20] Singh, M. and Fang, Y., ”Emotion Recognition in Audio and Video Using Deep Neural Networks”, arXiv preprint arXiv:2006.08129, 2020.

[21] Jia, N., and Zheng, C. and Sun, Wei., ”A Multimodal Emotion Recognition Model Integrating Speech, Video and MoCAP”, Multimedia Tools and Applications, 81(22): 32265–32286, 2022.

[22] Ren, M., Nie, W., Liu, A. and Su, Y., ”Multi-modal Correlated Network for Emotion Recognition in Speech”, Visual Informatics, 3(3): 150-155, 2019.

[23] Zheng, J., Zhang, S., Wang, Z., Wang, X. and Zeng, Z., ”Multi-channel Weight-sharing Autoencoder based on Cascade Multi-head Attention for Multimodal Emotion Recognition”, IEEE Transactions on Multimedia, 2022.

[24] Sharafi, M., Yazdchi, M., Rasti, R. and Nasimi, F., ”A Novel Spatio-temporal Convolutional Neural Framework for Multi-modal Emotion Recognition”, Biomedical Signal Processing and Control, 78, 2022.

[25] Chang, X. and Skarbek, Władysław., ”Multi-modal Residual Perceptron Network for Audio-video Emotion Recognition”, Sensors, 21(16): 5452, 2021.

Masker Language and Acoustic Confusability: Effects on Letter Sequence Recognition

Jessica L. L. Chin, Laurence Bruggeman, Mark Antoniou

The MARCS Institute for Brain, Behaviour, and Development, Western Sydney University

jessica.chin/l.bruggeman/m.antoniou@westernsydney.edu.au

Abstract

Speech-in-speech recognition is harder when the target and masker language are the same, and when cognitive load increases, but it is unclear whether the listener's language experience and the masker's similarity to the target also have an effect. Here, English monolinguals and Arabic–English bilinguals recalled English letter sequences with low or high acoustic confusability and embedded in masker speech. Results showed that a masker in the language of the target speech and high confusability are most detrimental to speech recognition. The Swedish masker's similarity to English did not affect recognition, nor did the bilinguals' familiarity to the Arabic masker.

Index Terms: speech-in-speech recognition, linguistic release from masking, linguistic similarity hypothesis, cognitive load

1. Introduction

Listeners can understand speech even in suboptimal conditions, such as noisy environments. They attend to a talker's speech (the target) whilst inhibiting distracting speech or other sounds (the masker) in a phenomenon commonly known as the "cocktail party effect" [1], [2]. If the masker becomes too distracting, recognising the target speech becomes hard or even impossible. We distinguish two main types of masking: energetic and informational. During energetic masking, the spectral qualities of the masker signal inhibit the intelligibility of the target speech. Informational masking refers to the additional aspects of a signal (outside of energetic masking) that can hinder target recognition, such as linguistic features in masker speech [1]. For instance, when listening to an English target, an English masker might make recognition more difficult than a Dutch masker due to the maskers' linguistic differences.

Some common findings arise in speech-in-speech recognition research. For instance, the more typologically similar a masker is to the target, the more difficult speech recognition becomes. This has been observed in multiple studies where the target and masker are the same language, e.g., English-in-English [3]–[6]. However, results have been varied for conditions where the target and masker languages differ, but are typologically similar. While maskers with the same rhythmic structure as the target (e.g., stress-, syllable-, or mora-timed) have been shown to be more detrimental to speech recognition than rhythmically dissimilar maskers [7], a recent study showed no significant differences in speech recognition accuracy between rhythmically similar or dissimilar maskers [8]. Indeed, rhythmic structure is only one element in which languages may share similarities, and similarities in other typological features would be worth exploring.

When accounting for the listener's language experience and their familiarity to a masker in speech-in-speech recognition, results have also been varied. Some studies have found that nonnative (L2) listeners are poorer at recognising speech within speech than native (L1) listeners [9], [10]. When the masker language is known to the L2 listeners but not the L1 listeners (e.g., English-in-Mandarin), L2 listeners perform worse in this condition [11]. This suggests that listeners encounter difficulties when they know the masker language, even when the target and masker language are mismatched. When comparing between more proficient listeners (i.e., late and early bilinguals), speech-in-speech recognition performance also varies. For instance, Spanish–English bilinguals who acquired English past age 14 found it harder to recognise English-in-English speech than English monolinguals and Spanish–English bilinguals who acquired English before age 6 [12]. Another study, however, showed that early Spanish–English bilinguals performed worse in English sentence recognition than monolinguals [13]. In contrast to the above findings, a study comparing English monolinguals and English–Greek bilinguals show comparable results between both groups for English targets in a Greek masker: a masker native to the bilinguals, yet foreign to the monolinguals [14]. In sum, a listener's language experience and their knowledge of a masker (that is not the same as the target language) can influence their speech-in-speech recognition performance, but there is no consensus for these patterns within the literature.

Speech-in-speech recognition typically requires more cognitive resources than listening to speech in quiet, as listening to a target while trying to ignore a masker requires attentional effort [15]. When considering the mechanisms involved in processing speech, the phonological loop allows for the short-term retention of verbal information, which also includes novel speech input [16]. There is also a link between acoustic representations of information and memory span, in which even visual representations of acoustically confusable letter sequences (e.g., BCD…) may lead to poorer retention scores than acoustically different (e.g., QYZ…) sequences [17]. Similar to the dividing of attention between multiple speech streams, the demand on working memory arising from implementing a secondary task can also make speech recognition more difficult [15]. In the current study, cognitive load, in the form of acoustical confusability of the target letter sequence, was manipulated to examine its effects on speech-in-speech recognition alongside masker language and the language experience of the listener.

The present study investigated the different linguistic and cognitive factors which influence speech-in-speech recognition. We aimed to determine how the phonetic similarity of the target and the masker languages affect speech-in-speech recognition. In addition, we were interested in whether the listener's knowledge of a masker language also impacts on speech-in-speech recognition. Not much is known about how these factors interact with each other. Also unclear is whether these factors impact speech recognition to the same extent, or whether one factor perhaps has a greater impact than the other. The task we used was letter sequence recognition in masker speech. We also manipulated the cognitive load—by using sequences with either low or high acoustic confusability—to assess any further impacts on the listener's speech-in-speech recognition performance. This replicates a classic paradigm from Conrad and Hull [17], in which more acoustically confusable letter sequences were recalled with lower accuracy. These sequences, however, were presented visually. Our current study instead presents auditory letter sequences with varying levels of acoustic confusability, all in the presence of masker speech.

In sum, our research question is as follows: How is speech-in-speech recognition influenced by: 1) the phonetic similarity between the target and masker languages, 2) the listener's familiarity to the masker language, and 3) the cognitive load demands of the task?

To answer this question, we compared two groups of participants (Australian English monolinguals, Arabic–English bilinguals) on their performance recalling an auditory sequence of English letters. The sequences were presented within two-talker masker speech in one of four masker languages (Australian English, Arabic, Swedish, Spanish). Masker languages varied in their similarity to English (as measured by the size of their vowel inventory), as well as in their familiarity to the listener.

We predicted that the English masker would be most detrimental, as it is the same language as the target, and also the native language of all participants. We also predicted that the Arabic masker would pose more difficulty for the bilinguals than the monolinguals. If native knowledge of a masker is more detrimental to performance than phonetic similarity between the masker and target, we predict the following patterns: for the monolinguals, the English masker would be the most difficult condition, followed by Swedish, then Spanish or Arabic. For the bilinguals, English would also be the most difficult condition, followed by Arabic, then Swedish, then Spanish. Finally, we predicted that more acoustically confusable letter sequences would be more difficult to recall than acoustically different ones.

2. Method

2.1. Participants

Participants were 61 Australian English (AusE) monolinguals (43 females, 1 non-binary individual; M_{Age} = 25.73, *SD* = 7.95) and 20 Arabic–English bilinguals (17 females; M_{Age} = 21.97, *SD* = 7.18). Data from two bilingual participants were discarded for failure to follow the task instructions. Participants were psychology undergraduates from Western Sydney University and were reimbursed with credit towards their course. All included participants signed informed consent. None reported any hearing or vision impairments, nor any learning or language disorders.

All participants completed a demographics questionnaire modelled after the Language Experience and Proficiency Questionnaire (LEAP-Q, [18]). The AusE monolinguals reported knowledge of only English and were all born in Australia, except for one participant who moved to Australia at the age of 4. The Arabic–English bilinguals acquired both languages within the first 10 years of life: four acquired both English and Arabic simultaneously, 10 acquired Arabic first, and five acquired English first. In this group, 10 participants were born in Australia, and three arrived in Australia before the age of 5. In the questionnaire, participants also rated their proficiency in reading, writing, speaking, and listening in each of their language(s) on a 9-point Likert scale, with 1 denoting no proficiency, and 9 native proficiency. The bilinguals reported an average score of 5 or greater across the domains of reading, writing, speaking and listening proficiency in English. In Arabic, they also reported an average score of 5 or greater for speaking and listening only, while reading and writing proficiency varied. However, we decided this was acceptable given that only spoken Arabic was presented in the experiment. The bilinguals identified with various cultures, including Syrian, Egyptian, Lebanese, Palestinian, Jordanian, and Saudi. One bilingual participant also spoke French but reported an average score of below 5 for their reading, writing, speaking, and listening proficiency.

2.2. Stimulus materials

Target stimuli were sequences of letters from the English alphabet spoken by a female Australian English speaker. The letters were recorded individually in a sound-attenuated booth at a sample rate of 44.1 kHz (16-bit). The letters were then grouped by cognitive load condition, randomised, and concatenated into a five-letter sequence using Praat [19]. Each letter was presented at an interstimulus interval of 250 ms, and stimuli were normalised to 65 dB sound pressure level (SPL). In the low cognitive load condition, the letters were acoustically dissimilar to one another (H, J, K, Q, Y, Z). In the high cognitive load condition, the letters either shared the same onset phoneme /e/ (F, L, M, N, S, X), or the same offset phoneme /iː/ (B, C, D, G, P, T, V), making them acoustically similar.

For the masker conditions, two female native speakers of each language (Australian English, Arabic, Swedish, and Spanish), for a total of eight talkers, produced 336 sentences from the Syntactically Normal Sentence Test list (SNST; [20]). These sentences are controlled for phrase structure and word frequency, and are semantically anomalous (e.g., "The salt dog caused the shoe"), since semantically meaningful masker speech makes speech-in-speech recognition even harder [3]. Like Australian English, which contains 19 vowels (monophthongs and diphthongs), Swedish contains a large number of vowels at 21 [21]. Conversely, Arabic only contains 8 [22], and Spanish 5 [23]. We first created a single masker track per talker by concatenating all of that talker's sentences in a random order. Using a script from Brouwer [24], the long-term average speech spectra for all masker tracks were normalised in Praat to minimise any spectral attributes which could interfere with the linguistic effects of masking. The masker tracks were then normalised to 70 dB SPL.

To create the final stimuli, each letter sequence was then combined with masker speech from both talkers of a language, excised from each masker track at a random starting point. The masker started 500 ms before the onset of the target letter sequence and ended 500 ms after target offset, in order to emulate the continuous stream of background speech encountered in natural listening environments. The final stimuli were presented at a signal-to-noise ratio (SNR) of -5 dB.

2.3. Procedure

The experiment was conducted remotely using E-Prime Go 1.0 [25]. Participants were instructed to sit at a table in a quiet environment and used their own computer and headphones. The task started with four practice trials, presented in quiet (two trials) or +10dB SNR (two trials). Participants then completed the 80 experiment trials (4 maskers × 20 targets) at -5 dB SNR, with half of all trials per masker condition in the low cognitive load condition, and the other half in the high cognitive load condition. Within the high cognitive load condition, 5 trials contained letters sharing an identical onset phoneme, and 5 shared an identical offset phoneme. All conditions were presented in a mixed, randomised order across the 80 test trials.

In each trial, participants heard a sequence of five English letters presented in masker speech. After a 1 second delay from the stimulus offset, participants were prompted to type the letters they heard in the order that they heard them. After pressing the ENTER key to submit their response, they then proceeded to the next letter sequence.

3. Results

Data from one AusE monolingual were excluded from analysis, as this participant's mean accuracy in the Arabic, Spanish, and Swedish masker conditions was 3 standard deviations from the group mean. The analyses below thus include data from 60 AusE monolingual participants.

The dependent variable was participants' letter recognition accuracy per trial, measured as the Levenshtein edit distance, which is the minimum number of insertions, deletions, and/or substitutions required to match two strings (i.e., the response and the presented letter sequence) [26]. A lower Levenshtein distance indicates higher similarity between two strings, and thus a better recognition of the target sequence. For the target sequence HJKQY, a response of HJKQY would have a Levenshtein distance of 0 (no edits required), a response of HJKQM would result in a distance of 1 (substitution of the final letter), while ABCDE would have a distance of 5 (all characters substituted). The Levenshtein distance was calculated for each trial using the *stringdist* package [27] in R Version 4.4.1 [28]. For consistency, only the first 5 letters in each response were considered. Mean Levenshtein distances for the AusE monolingual and Arabic–English bilinguals across masker and cognitive load conditions are shown in Figure 1.

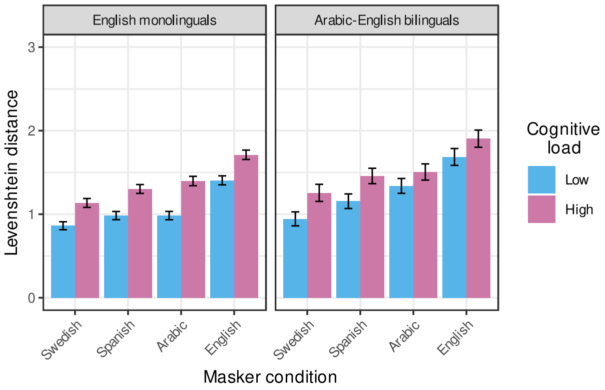


Figure 1: *Mean Levenshtein distance between target sequence and response by listener group across masker and cognitive load. Error bars depict the standard error.*

Using the *brms* package [29] in R, a Bayesian generalised linear mixed model was fitted to the data, with the treatment-coded fixed effects Listener Group (AusE [reference level], Arabic–English), Masker (English [reference level], Arabic, Swedish, Spanish), and Cognitive Load (Low [reference level], High). To determine whether there was a learning effect, the Trial Number variable was standardised to a mean of 0 and standard deviation of 1 using R's *scale* function, then included as a fixed factor. The model also included the random intercept for item and by-participant random slopes for Listener Group. We used weakly informative priors (i.e., a normal distribution with a mean of 0 and a standard deviation of 1), and a binomial distribution.

Directional hypothesis tests were run based on the following predictions. For both listener groups, a higher Levenshtein distance (i.e., poorer letter sequence recognition) would be observed for the English masker condition (identical language to target) compared to the other maskers. For the monolinguals, the Swedish masker would have the second highest Levenshtein distance (phonetically similar to target).

Table 1. *Directional hypothesis tests for the effects of Listener Group, Masker, and Cognitive Load. Mean = mean posterior distribution, 90% CI = one-sided 90% credibility interval, ER = evidence ratio, PP = posterior probability. As the dependent variable is inversely related to performance, a hypothesis of the form* a > b *means that* a *is hypothesised to affect performance more negatively than* b.

Hypothesis	Mean	90% CI	ER	PP
AusE group				
Swedish < English	-0.67	[-0.84, -0.50]	> 7999.00	1.00
Spanish < English	-0.48	[-0.65, -0.31]	> 7999.00	1.00
Arabic < English	-0.43	[-0.60, -0.26]	> 7999.00	1.00
Spanish < Swedish	0.19	[0.02, 0.37]	0.03	0.03
Arabic < Swedish	0.24	[0.07, 0.41]	0.01	0.01
Arabic–English group				
Swedish < English	-0.76	[-0.97, -0.56]	> 7999.00	1.00
Spanish < English	-0.51	[-0.72, -0.31]	7999.00	1.00
Arabic < English	-0.39	[-0.59, -0.19]	1332.33	1.00
Spanish < Swedish	0.06	[-0.10, 0.21]	0.38	0.27
Arabic > Swedish	0.13	[-0.03, 0.28]	10.41	0.91
Arabic > Spanish	0.07	[-0.08, 0.22]	3.73	0.79
Both groups				
High cog. load > low cog. load	0.37	[0.25, 0.48]	> 7999.00	1.00
Scaled trial < 0	0.02	[-0.01, 0.04]	0.13	0.12
AusE < Arabic–English	0.25	[-0.15, 0.64]	5.81	0.85

For the bilinguals, the Arabic masker would have the second highest Levenshtein distance (listeners' native language), followed by the Swedish masker. Finally, for both groups, the high cognitive load condition (more acoustically confusable letter sequences) would result in a higher Levenshtein distance than the low cognitive load condition.

Evidence ratios (ER), which is the ratio of the posterior probability (PP) of the test hypothesis and that of the alternate hypothesis, were used to determine whether there was strong evidence for the tested hypotheses. An ER of over 19 is analogous to a p-value of < 0.05 in frequentist statistics, and is considered strong evidence in support of the test hypothesis [30].

Hypothesis test results (see Table 1) show that, for the AusE monolinguals, there is extremely strong evidence that the Swedish, Spanish, and Arabic maskers were less detrimental to performance than the English masker, with PPs of 1.00 and ERs of infinity (*inf*). A PP of 1.00 indicates that 100% of the posterior samples fall on the side of the test hypothesis, while an ER of infinity can be read as greater than S - 1, where S is the number of posterior draws used in the model (i.e., 8000, giving an ER of > 7999). For the Arabic–English bilinguals; the Swedish, Spanish, and Arabic maskers were also less detrimental to performance than English maskers. There was also very strong evidence that the high cognitive load condition resulted in higher Levenshtein distance scores than the low cognitive load condition.

However, there was no evidence that a phonetically similar non-English masker (Swedish) was more detrimental to letter sequence recognition than a dissimilar masker (Arabic and Spanish); this was the case for both AusE monolinguals and Arabic–English bilinguals. For the Arabic–English bilinguals, the Arabic masker also was not more detrimental to letter sequence recognition than Swedish or Spanish. The listener groups also did not differ in performance across masker conditions. Furthermore, there was no evidence for any learning effects over the course of the task.

4. Discussion

The present study investigated the effects of listener group, masker language, and cognitive load condition during speech-in-speech recognition, using a letter sequence recall task. As predicted, the findings suggest that letter sequence recognition is more difficult when the target and masker language are the same. This is consistent with findings from previous studies on this topic [3]–[5]. We also found that recognition is poorer when the letters in a sequence are acoustically confusable, in line with past studies that also observed this phenomenon, albeit with written stimuli [17]. It should be noted that data collection is ongoing, so a larger sample size for the Arabic–English bilinguals may uncover additional speech-in-speech recognition patterns regarding listeners' familiarity to the masker speech.

Our predictions regarding the phonetic similarity of the non-English maskers were not borne out. The Swedish masker does not appear to hinder letter sequence recognition more than Spanish or Arabic, even though Swedish is more similar to English both phonetically and rhythmically (both languages are stress-timed [21]). In fact, it appears that the Swedish condition was easier to complete than the other maskers. Possible explanations arise for why the Swedish masker was less detrimental to speech-in-speech recognition performance. Swedish, as a pitch-accented language [21], may sound less speechlike and more "musical" to our participants due to its tonal characteristics, allowing our participants to more easily ignore it while attending to the target speech. On a similar note, it is possibly less commonly-heard than Arabic and Spanish by our participants, who reside in Australia. When participants were asked after the task which languages they had heard, none identified Swedish as one of the maskers, but several were able to identify Spanish and Arabic. Future studies can examine these points further, for instance, by using German (which shares the same rhythmic structure as English and a large vowel inventory) as a masker language. Listeners in Australia are also likely to be more exposed to German as a foreign language than Swedish or other Germanic languages. In addition, a study examining various pitch accented and/or lexical tone masker languages could be an avenue to explore whether tonality in languages contributes to how speechlike a masker is perceived.

We also found no evidence that speech-in-speech recognition is inhibited by native knowledge of a masker (when it is *not* the target language). For the Arabic–English bilinguals, listening to an Arabic masker had no negative impact on their performance, even when compared to their monolingual counterparts. In the literature, reports of the impact of listening to a masker known to the listener are varied; poorer speech-in-speech recognition accuracy is more consistently observed in nonnative listeners [9]. The current findings align with studies which have found that monolinguals and bilinguals perform comparably in speech-in-speech recognition, even when the masker is the other language spoken by the bilinguals [14]. The Arabic–English bilinguals in this study were also early bilinguals, and despite currently being English dominant, reported high levels of Arabic speaking and listening proficiency. Furthermore, it should be noted that the variety of Arabic used for our maskers was Modern Standard Arabic. We selected this variety so that participant recruitment would not be restricted to listeners of a single regional variety of Arabic, although it meant that the masker dialect was not native to any of our bilingual participants. While all bilingual listeners correctly identified Arabic as one of the masker languages upon completion of the experiment, this may nevertheless have affected the degree to which they were hindered by the Arabic masker. Future speech-in-speech recognition research could compare the effects of a listener's familiarity to their native regional variety of Arabic versus the prestige Modern Standard Arabic variety.

In sum, speech-in-speech recognition was most difficult when the target and masker language were identical, as well as when the letter sequence was more acoustically confusable. Neither a phonetically-similar masker (i.e., Swedish) nor a non-English masker known to the bilingual listeners (i.e., Arabic) were shown to negatively impact speech-in-speech recognition in this study. Regardless, there is potential for future research into speech-in-speech recognition focusing on suprasegmental qualities of masker languages, such as pitch, as well as the listener's familiarity to dialectal varieties of a language.

5. Acknowledgements

We would like to thank Jade Nguyen, Gabrielle Teklic, Bishoy Boulos, Elena Talevska, Adau Aher, Ella Ward, Jasmine Abdel Wahab, Kristina Petkovich, Afrah Parkar, Mai Linh Tran, and Umair Yakub for their contributions to experimental stimuli development and participant recruitment.

6. References

[1] Kidd, G. and Colburn, H. S., “Informational masking in speech recognition”, in J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay, [Eds.], The Auditory System at the Cocktail Party, 60:75–109, Springer International Publishing, 2017.

[2] Pollack, I., “Auditory informational masking”, J. Acoust. Soc. Am., 57:S5–S5, 1975.

[3] Brouwer, S., Van Engen, K. J., Calandruccio, L. and Bradlow, A. R., “Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content”, J. Acoust. Soc. Am., 131:1449–1464, 2012.

[4] Garcia Lecumberri, M. L. and Cooke, M., “Effect of masker type on native and non-native consonant perception in noise”, J. Acoust. Soc. Am., 119:2445–2454, 2006.

[5] Van Engen, K. J. and Bradlow, A. R., “Sentence recognition in native- and foreign-language multi-talker background noise”, J. Acoust. Soc. Am., 121:519–526, 2007.

[6] Williams, B. T. and Viswanathan, N., “The effects of target-masker sex mismatch on linguistic release from masking”, J. Acoust. Soc. Am., 148:2006–2014, 2020.

[7] Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S. and Bradlow, A. R., “Masking release due to linguistic and phonetic dissimilarity between the target and masker speech”, Am. J. Audiol., 22:157–164, 2013.

[8] Brown, V. A. et al., “Revisiting the target-masker linguistic similarity hypothesis,” Atten. Percept. Psychophys., 84:1772-1787, 2022.

[9] Garcia Lecumberri, M. L., Cooke, M. and Cutler, A., “Non-native speech perception in adverse conditions: A review”, Speech Commun., 52:864–886, 2010.

[10] Cooke, M., Garcia Lecumberri, M. L. and Barker, J., “The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception”, J. Acoust. Soc. Am., 123:414–427, 2008.

[11] Van Engen, K. J., “Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble”. Speech Commun., 52:943–953, 2010.

[12] Mayo, L. H., Florentine, M. and Buus, S., “Age of second-language acquisition and perception of speech in noise”, J. Speech Lang. Hear. Res., 40:686–693, 1997.

[13] Krizman, J., Bradlow, A. R., Lam, S. S.-Y. and Kraus, N., “How bilinguals listen in noise: Linguistic and non-linguistic factors”, Biling. Lang. Cogn., 20:834–843, 2017.

[14] Calandruccio, L. and Zhou, H., “Increase in speech recognition due to linguistic mismatch between target and masker speech: Monolingual and simultaneous bilingual performance”, J. Speech Lang. Hear. Res., 57:1089–1097, 2014.

[15] Mattys, S. L., Davis, M. H., Bradlow, A. R. and Scott, S. K., “Speech recognition in adverse conditions: A review”, Lang. Cogn. Process., 27:953–978, 2012.

[16] Baddeley, A., “Working memory: looking back and looking forward,” Nat. Rev. Neurosci., 4:829–839, 2003.

[17] Conrad, R. and Hull, A. J., “Information, acoustic confusion and memory span”, Br. J. Psychol., 55:429–432, 1964.

[18] Marian, V., Blumenfeld, H. K. and Kaushanskaya, M., “The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals”, J. Speech Lang. Hear. Res., 50:940–967, 2007.

[19] Boersma, P. and Weenink, D., “Praat: Doing phonetics by computer”, 2022.

[20] Nye, P. W. and Gaitenby, J. H., “The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences", in Haskins Laboratories Status Report on Speech Resolution, SR-37/38, 169–190, 1974.

[21] Riad, T., The Phonology of Swedish, Oxford University Press, 2014.

[22] Watson, J. C. E., The Phonology and Morphology of Arabic, OUP Oxford, 2007.

[23] Green, J. N., “Spanish”, in M. Harris and N. Vincent, [Eds.], The Romance Languages, 79–130, Taylor & Francis, 1997.

[24] Brouwer, S., “The role of foreign accent and short-term exposure in speech-in-speech recognition”, Atten. Percept. Psychophys., 81:2053–2062, 2019.

[25] Psychology Software Tools, Inc., “E-Prime Go.” 2020.

[26] Levenshtein, V. I., “Binary codes capable of correcting deletions, insertions, and reversals”, Sov. Phys. Dokl., 10:707–710, 1966.

[27] van der Loo, M. P. J., “The stringdist package for approximate string matching”, R J., 6:111–122, 2014.

[28] R Core Team, “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, 2024.

[29] Bürkner, P.-C., “brms: An R package for Bayesian multilevel models using Stan”, J. Stat. Softw., 80:1–28, 2017.

[30] Milne, A. J. and Herff, S. A., “The perceptual relevance of balance, evenness, and entropy in musical rhythms”, Cognition, 203:104233, 2020.

Coarticulation and Transitional Probability in Japanese Perceptual Epenthesis

Alexander J. Kilpatrick[1], Sharon Peperkamp[2]

[1]Nagoya University of Commerce and Business, [2]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS-PSL, CNRS, EHESS

alexander_kilpatrick@nucba.ac.jp, sharon.peperkamp@ens.psl.eu

Abstract

This study examines the roles of coarticulation and transitional probability in Japanese perceptual epenthesis, where listeners perceive illusory vowels to resolve phonotactically illegal sequences. We replicated and extended a previous experiment showing that Japanese listeners tend to insert /u/ or /i/ in illicit sequences like /eʃpo/ and /et͡ʃpo/, influenced by the transitional probability of the preceding consonant. To isolate the effects of coarticulation, we modified /et͡ʃpo/ stimuli by removing the closure of the affricate /t͡ʃ/. Participants categorized and discriminated between these stimuli in two experiments. Results indicated that even with reduced coarticulatory influence, the patterns of /u/ and /i/ epenthesis persisted, underscoring the significant role of transitional probability in perceptual epenthesis. This study clarifies how Japanese listeners utilize both coarticulation and transitional probability to perceptually adapt illicit consonant clusters into phonotactically licit sequences.

Index Terms: perceptual epenthesis; Japanese; speech perception; transitional probability; coarticulation

1. Introduction

Perceptual epenthesis is a phenomenon where listeners perceive illusory vowels within sequences that violate the phonotactic rules of their native language [1,2]. This process is a perceptual strategy to resolve illicit sequences into phonotactically legal sequences. For instance, Japanese listeners often hear an epenthetic /u/ in sequences like /ebzo/, perceptually transforming it into /ebuzo/ to conform to Japanese phonotactics which typically prohibits non-homorganic consonant clusters [1]. This perceptual adjustment is not unique to Japanese but is observed across various languages, each with its own tendencies for epenthetic vowel selection. In Japanese, the most common epenthetic vowel is /u/, which is phonetically minimal due to its short duration and tendency to undergo devoicing [3].

In this study, we replicate and extend an earlier experiment on Japanese perceptual epenthesis [4], which showed that listeners epenthesized /u/ or /i/ depending on the preceding consonant in phonotactically illicit tokens of /eʃpo/ and /et͡ʃpo/. This was argued to be the result of the relative transitional probability of CV sequences. Transitional probability is the likelihood of one sound following another which in turn is used to calculate Surprisal and Entropy. While both /t͡ʃi/ and /ʃi/ are reasonably high-frequency diphones (low Surprisal [5]), the context following /t͡ʃ/ is less chaotic (low Entropy [5]) than the context following /ʃ/ in Japanese, and the former was found to elicit a relatively high rate of non-default /i/-epenthesis. Yet, an alternative account may be that the coarticulatory influence of the preceding vowel /e/ enhances /i/-epenthesis, as demonstrated in [6]. This coarticulation might be obstructed by the closure in the affricate /t͡ʃ/, potentially affecting the selection of the epenthetic vowel. To investigate this possibility, we modified the /et͡ʃpo/ stimuli by removing the closure at the beginning of the affricate, which should reduce the coarticulatory influence of the preceding vowel. In the editing process, we took an average of the unmodified fricatives' duration as a reference and aimed to mirror that average by extending the frication in the modified stimuli, ensuring that the editing process did not present perceptible signs of alteration. We then ran the same experiment with the inclusion of the modified stimuli to determine whether the observed patterns of /u/ and /i/ epenthesis persist under these modified conditions.

1.1. Epenthetic Vowels

The choice of epenthetic vowel can differ significantly between languages, influenced by a range of phonological, statistical, and acoustic factors. For example, in Brazilian Portuguese, the shortest vowel, /i/, functions as the default epenthetic segment, contrasting with the /u/ in Japanese [2]. In Korean, the dominant epenthetic vowel is /ɨ/; however, listeners may perceive an alternate epenthetic vowel /i/ in sequences where the production of the dominant vowel would violate Korean phonotactics [7]. Spanish listeners have been found to insert an /e/ in sequences that disallow consonant clusters [8], and English speakers may use the schwa (/ə/) [9].

A seemingly important feature of epenthetic vowels is that the target vowel or vowels can sometimes be omitted without changing the meaning of words. For example, in English the centralized schwa is both the target for vowel reduction and perceptual epenthesis. The different ways that the word *family* can be pronounced illustrates this concept because whether one says [fæmɪli], [fæməli], or [fæmli], the meaning of the word remains unchanged. In Japanese, high vowels (/i/ and /u/) can undergo a process known as vowel devoicing under certain phonological conditions. This devoicing occurs when high vowels are situated between voiceless consonants or at the end of a word following a voiceless consonant. Despite the devoicing, and sometimes even deletion (see [3]), these vowels are still considered to be part of the underlying phonological representation of the word. As a result, the meaning of the word remains unchanged even when the high vowel is not audibly pronounced. This phenomenon is a part of the native speaker's phonological knowledge and does not typically lead to confusion or ambiguity in understanding the spoken language.

Another seemingly important feature in epenthetic vowel selection is coarticulation, where phonotactically illegal consonant clusters are affected by acoustic cues from adjacent vowels. For example, [6] used an identification task to assess the quality of epenthetic vowels perceived by Japanese listeners in illegal consonant clusters with varying amounts of coarticulation. They created cross-spliced stimuli where the coarticulation present in the consonant cluster did not match the quality of the flanking vowel. Two types of clusters were used: /hp/ and /kp/, with the former containing larger amounts of resonances from the preceding vowel. The results showed that both the flanking vowel and coarticulation influenced the quality of the perceived epenthetic vowel, but the influence of coarticulation was much larger for /hp/ clusters because the vowel identity was better signaled through the fricative in /hp/ clusters compared to the stop in /kp/ clusters.

Transitional probability also seems important in the selection of epenthetic vowels during perceptual epenthesis. This concept refers to the likelihood of a particular vowel following a given consonant based on a listener's linguistic experience. [4] found that Japanese listeners were more likely to perceive an epenthetic /i/ following the consonant clusters /ʃ/ and /t͡ʃ/ due to the higher transitional probability associated with these combinations as opposed to /ɡ/. Transitional probability was calculated using the Surprisal and Entropy [5] equations, where Surprisal indicates how unexpected a vowel is after a consonant, and Entropy reflects the uncertainty or variability of which vowel might follow. For instance, the Surprisal values for /ʃi/ and /t͡ʃi/ are lower than those for /ʃu/ and /t͡ʃu/, meaning that /i/ is less surprising and more predictable after /ʃ/ and /t͡ʃ/ than /u/ is. Additionally, the Entropy for /t͡ʃ/ is lower than for /ʃ/, indicating that the vowel following /t͡ʃ/ is more predictable. The results showed that Japanese listeners were more likely to experience /i/ epenthesis following /t͡ʃ/ than following /ʃ/ despite both having a high transitional probability (or low Surprisal) with /i/. In other words, despite /u/ being the default target for perceptual epenthesis in Japanese, both /t͡ʃ/ and /ʃ/ elicited /i/ epenthesis with a slightly stronger effect for /t͡ʃ/ that reflects both the transitional probability of /t͡ʃi/ (Surprisal), and the overall probability of the context following /t͡ʃ/ (Entropy).

In this study, we recreate the experiments featured in [4] with the inclusion of modified stimuli where those stimuli have been altered to minimize the coarticulatory influence of the preceding vowel. Specifically, we removed the closure at the beginning of the affricate /t͡ʃ/ in /et͡ʃpo/, /et͡ʃipo/, and /et͡ʃupo/ stimuli which should reduce the carryover effects of coarticulation and allow us to better isolate the role of transitional probability in epenthetic vowel selection. By comparing the patterns of /u/ and /i/ epenthesis in both the original and modified stimuli, we aim to determine whether the previously observed tendencies are robust to changes in coarticulatory context. This approach allows us to disentangle the relative contributions of coarticulation and transitional probability to the perceptual epenthesis process, providing a clearer understanding of how Japanese listeners resolve phonotactically illegal sequences in their native language.

2. Method

All models were constructed in R [10]. All data and code relating to this project can be found here: https://osf.io/mduc9/?view_only=e20f9f4d92d54044b0ab941122b3d24a. The following experiments have ethics approval from the Nagoya University of Commerce and Business (#23065).

2.1. Stimuli

The stimuli for both experiments are shown in Table 1. Other than the modified tokens, these are the exact same stimuli as used in [4]. The medial /C/ and /CV/ sequences were embedded in in a nonce frame constructed from /e/, /p/, and /o/ because these phonemes occur infrequently in Japanese and thus limit unintended variability to the dataset beyond those factors deliberately manipulated. Tokens were constructed into either licit /eCVpo/ or illicit /eCpo/ sequences. Vowels presented in the epenthetic position in licit sequences were either /i/ or /u/ and consonants preceding the epenthetic position were /ʃ/ or /t͡ʃ/. Three female Australian English speakers produced five tokens of each target with initial stress. The first and fifth tokens were discarded. Recordings were conducted in the Horwood recording studio at the University of Melbourne and were recorded in mono with 16-bit resolution.

Table 1. *List of stimuli used in both experiments. Modified affricates are explained in the following paragraph.*

Token	/eCpo/	/eCupo/	/eCipo/
Fricative	/eʃpo/	/eʃupo/	/eʃipo/
Affricate	/et͡ʃpo/	/et͡ʃupo/	/et͡ʃipo/
Modified Affricate	/e(t)͡ʃpo/	/e(t)͡ʃupo/	/e(t)͡ʃipo/

All tokens with the affricate /t͡ʃ/ were manipulated in Praat. To edit the /t͡ʃ/ tokens to resemble /ʃ/ tokens (represented with /(t)͡ʃ/), we first identified and spliced out the closure portion of the affricate (highlighted in Figure 1), from near the beginning of the drop in energy after the /e/ to just after the release in /t͡ʃ/. Then, we extended the remaining aperiodic portion by splicing a section of it from the middle of the aperiodic period into the existing fricative to mask any acoustic cues of the edit, and to emulate the longer duration characteristic of the [ʃ] in /ʃ/ tokens. Each token was carefully analyzed to ensure that the editing process did not introduce any additional pops or other indication of token manipulation. Although considerable care was taken in this process, the average unmodified fricative length (*M* = 118ms, *SD* = 23) was slightly longer than the average modified fricative length (*M* = 114ms, *SD* = 22).

2.2. Participants

We recruited 44 Japanese undergraduate students from the Nagoya University of Commerce and Business started the experiment. All participants reported being monolingual Japanese speakers. Two participants requested to end the experiment prior to completion and their data was discarded. Of the 42 participants that completed the experiment, 9 students reported having at least one parent with a native language other than Japanese. These include 4 Brazilian Portuguese, and one each Peruvian Spanish, Turkish, English, Urdu, and Mandarin Chinese. Given that /i/ is the default epenthetic vowel in Brazilian Portuguese [2], we eyeballed the data from those participants with Brazilian Portuguese speaking parents in the AXB discrimination experiment and found their results to be different to those from the other participants. We therefore exclude the results of these 4 participants unless otherwise noted. Participants signed consent forms prior to starting the experiment and were not compensated for their time.

2.3. Procedure

Both experiments were conducted using Psyscope X on iMacs running a 32bit operating system. Six computers were set up in a laboratory and the experiments were conducted on a maximum of six participants at a time. Computers were equipped with noise cancelling headphones, keyboards, and mice. Prior to the experiments, participants completed a short biographical survey. The experiments took less than 15 minutes to complete. The first author gave instructions in Japanese and waited for all participants to complete Experiment 1 prior to explaining the procedure for Experiment 2.

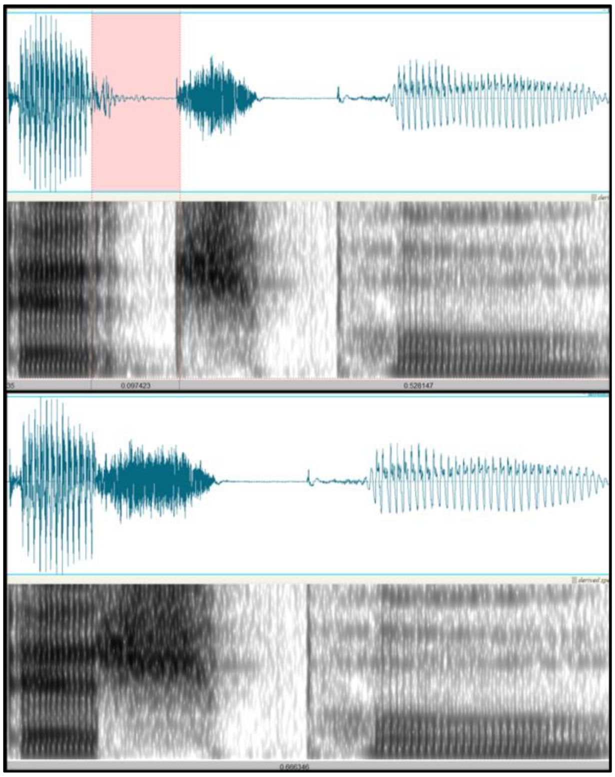


Figure 1: *Waveform and spectrogram of an /et͡ʃpo/ token with the highlighted closure segment that was removed to transform the affricate /t͡ʃ/ into the fricative /ʃ/ (top). Modified /e(t͡)ʃpo/ token after removal of the closure and extension of the fricative (bottom).*

2.3.1. *Experiment 1: Categorization*

In the categorization experiment, participants were exposed to the stimuli listed in Table 1. They were asked to categorize each string into one of four categories (/et͡ʃipo/, /et͡ʃupo/, /eʃipo/, and /eʃupo/), which were presented as on-screen buttons with hiragana labels (えちぽ, えちゅぽ, えしぽ, and えしゅぽ). Participant responses were recorded via mouse click. After assigning a category, the token would be replayed, and the participant would be shown a seven-point Likert scale. They were asked to assign a score according to how well the token fit the assigned category. The experimental strings were drawn randomly from a library of 81 tokens, with three repetitions of each of the nine stimuli from each of the three speakers. Each token was presented to participants one time unless they failed to respond within 3500ms, in which case the trial would be randomly inserted into the remainder of the experiment.

2.3.2. *Experiment 2: Discrimination*

In the discrimination experiment, participants were exposed to AXB triads made up of three tokens listed in Table 1. In this experiment, participants were asked to indicate whether the second token best matched the first or third token by pressing either "1" or "3" on a keyboard. To avoid sequence bias, the individual tokens for each AXB triad were counterbalanced and occurred in every position of the triad. Each triad comprised one token from each of the three speakers, determined by partial Latin square (123, 132, 213, 231, 312, 321, where the numbers denote a speaker), and each of these sequences was organized into four different triad sequences (AAB, ABB, BAA, BBA, where the letters denote a stimulus token). Each trial was drawn three times at random from this array of 24 triads, resulting in a total of 72 trials. Participants had a 2000 ms response window to respond. If they failed to respond within this time, the missed trial was replayed at a random time during the remainder of the experiment.

3. Results

3.1. Categorization

In the categorization experiment, the unmodified illicit tokens (/et͡ʃpo/ and /eʃpo/) behave similarly with both exhibiting around a 60/40 split with a preference for /u/ classification and a reasonably high average fit score (~6). As to the modified tokens, 24.2% of them were classified as /t͡ʃ/ tokens, suggesting that some elements of the affricate remain despite splicing out the closure and extending the fricative. The e(t͡)ʃpo stimuli were more often classified to the /eʃupo/ (53.24%) than the /eʃipo/ (21.18%) category. The fit scores for the modified stims were slightly lower (~5.5) than the unmodified stims.

Table 2. *Results of the Categorization experiment. Categories run across the top of the table; tokens are to the left. The first number represents the percentage of samples assigned to that category; the second number is the average fit score. Shaded cells reveal majority classification.*

		/et͡ʃipo/	/et͡ʃupo/	/eʃipo/	/eʃupo/
/eCipo/	et͡ʃipo	87.32% (6.17)	9.14% (4.90)	2.06% (4.71)	1.47% (3.80)
	eʃipo	0.58% (1.00)	0.29% (3.00)	92.42% (6.02)	6.71% (4.65)
	e(t͡)ʃipo	16.08% (5.75)	2.92% (3.00)	76.02% (5.67)	4.97% (5.59)
/eCupo/	et͡ʃupo	7.87% (5.93)	90.67% (5.83)	0.87% (5.00)	0.58% (4.00)
	eʃupo	0.58% (4.00)	2.04% (5.00)	2.04% (3.71)	95.34% (6.06)
	e(t͡)ʃupo	4.97% (5.53)	23.10% (5.66)	6.73% (4.96)	65.20% (5.62)
/eCpo/	et͡ʃpo	39.77% (5.93)	59.65% (5.97)	0.58% (2.50)	0%
	eʃpo	1.16% (5.75)	2.33% (3.25)	37.50% (5.67)	59.01% (5.99)
	e(t͡)ʃpo	10.88% (5.08)	14.71% (5.06)	21.18% (5.49)	53.24% (5.48)

3.2. Discrimination

Table 3. *Results of the Discrimination experiment. Acc% represents the discrimination accuracy in AXB trials.*

Contrast pair		Acc%
[et͡ʃpo]	[et͡ʃupo]	65.79%
	[et͡ʃipo]	63.82%
[eʃpo]	[eʃupo]	57.68%
	[eʃipo]	69.96%
[e(t)͡ʃpo]	[e(t)͡ʃupo]	57.02%
	[e(t)͡ʃipo]	64.91%

Using the *lme4* package [11], we analyzed these data in a logistic mixed-effects model with contrast-coded fixed factors Consonant (/t͡ʃ/ vs. /ʃ/ vs. /(t)͡ʃ/, Vowel (/i/ vs/ /u/) and their interaction, and a random intercept for Participant. The Anova function in the *Car* package [12] was used to establish statistical significance, and post-hoc analyses with corrections for multiple comparison were run in the *emmeans* package [13].

The model revealed an effect of Vowel ($\beta = 0.13$, $SE = 0.04$, $z = 3.31$, $\chi^2(2) = 11.0$, $p<.001$), with overall higher accuracy for the pairs with /i/ than for those with /u/, and a Consonant × Vowel interaction (/t͡ʃ/:/i/: $\beta = -0.18$, $SE = 0.06$, $z = -3.11$; /ʃ/:/i/: $\beta = 0.14$, $SE = 0.06$, $z = 2.46$; /(t)͡ʃ/:/i/: $\beta = 0.04$, $SE = 0.06$, $z = 0.65$; $\chi^2 = 10.7$, $p < .005$). Post-hoc analyses revealed that for the affricate accuracy did not differ according to the vowel ($\beta = -0.09$, $SE = 0.14$, $z < 1$), while for the fricative and the modified affricate accuracy was higher for pairs with /i/ than for those with /u/ (fricative: $\beta = 0.55$, $SE = 0.14$, $z = 3.90$; $p < .001$; modified affricate: $\beta = 0.34$, $SE = 0.14$, $z = 2.48$; $p < .02$). In addition, neither in the context of /i/ nor in that of /u/ was there a difference in accuracy between the fricative and the modified affricate (/i/: $\beta = 0.24$, $SE = 0.14$, $z = 1.65$; $p > .1$; /u/: $\beta = 0.03$, $SE = 0.14$, $z < 1$).

With a caveat of very small sample size, we also analyzed the possible influence of likely exposure to Brazilian Portuguese (BP), whose default epenthetic vowel is /i/ rather than /u/ [2]. Thus, we included the four participants with at least one native Brazilian Portuguese speaking parent in the sample and ran the same model as before but with the addition of contrast-coded BP (yes vs. no) and all its interactions. The mean accuracy scores of the added participants were 83.3% and 66.7% for the affricate when the vowel was /u/ or /i/, respectively, 77.1% and 72.9% for the fricative, and 68.8% and 72.8% for the modified affricate. The model revealed the same main effect of Vowel and its interaction with Consonant as found in the previous model. In addition, there was an effect of BP, with overall higher accuracy by the participants with a Brazilian Portuguese speaking parent ($\beta = 0.50$, $SE = 0.22$, $z = 2.26$, $\chi 2 = 5.10$, $p = 0.024$), and, crucially, a BP × Vowel interaction ($\beta = -0.29$, $SE = 0.14$, $z = -2.04$, $\chi^2 = 4.16$, $p < 0.05$). Post-hoc analyses revealed that contrary to the other participants, those with a Brazilian Portuguese speaking parent did not perform differently in the two vowel contexts ($\beta = -0.32$, $SE = 0.27$, $z = -1.15$, $p > 0.1$). In addition, in the context of /u/ the participants with a Brazilian Portuguese speaking parent had higher accuracy scores than the other ones ($\beta = -0.79$, $SE = 0.27$, $z = -2.94$, $p < .004$), while there was no difference in the context of /i/ ($\beta = -0.21$, $SE = 0.26$, $z < 1$).

4. Discussion

This study aimed to disentangle the relative contributions of vowel coarticulation and transitional probability to the perceptual epenthesis process by Japanese listeners in clusters starting with a fricative vs. an affricate. A previous study found that a cluster starting with /t͡ʃ/ yields more /i/-epenthesis than one starting with /ʃ/ [4]. By modifying the closure segment of the affricate /t͡ʃ/ in the stimuli used in that study, we created tokens that were predominantly classified as /ʃ/ stimuli. These modified tokens were then used together with the original tokens with /t͡ʃ/ and /ʃ/ in an AXB discrimination experiment. For the original tokens we qualitatively replicated the response pattern observed in [4], i.e. for the /ʃ/ tokens higher accuracy (hence less epenthesis) in the condition with /i/ than that with /u/, and no difference for the /t͡ʃ/ tokens. Crucially, even though as far as /i/-epenthesis is concerned the results for the modified tokens were numerically in between those for the original /ʃ/ and /t͡ʃ/ tokens, overall, the results for these tokens were statistically indistinct from those for the /ʃ/ tokens. These results confirm that transitional probability plays a significant role in epenthetic vowel quality. Thus, while coarticulatory cues are important [6], they are not the sole determinant in epenthesis patterns. Instead, transitional probability also influences the selection of the epenthetic vowel, contributing to the overall perceptual process.

The low classification accuracy of the pairs [eʃpo]-[eʃupo] and [e(t)͡ʃpo]-[e(t)͡ʃupo]—compared to other pairs—suggests that /u/ is a much stronger target for epenthesis in Japanese. This is not a particularly novel finding given that other studies [e.g., 2] have shown that Japanese listeners prefer /u/ in these contexts. Future studies might also include contrasts where unmodified tokens are tested against modified tokens, such as [e(t)ʃpo]-[eʃupo] and [e(t)ʃpo]-[eʃipo], to further explore the nuances of coarticulatory effects.

Finally, although we had not planned to investigate this issue, the presence of 4 participants with at least one native Brazilian Portuguese speaking parent allowed us to tentatively examine the influence of likely exposure to this language. That is, while these participants reported being monolingual, they had likely been exposed to BP to at least some extent from birth on. We found that these participants showed overall higher accuracy in the context of /u/ and hence perceived more /i/-epenthesis than the other participants. This finding is reminiscent of previous research showing that the default epenthetic vowel in second-generation Japanese immigrants in Brazil who have acquired both Japanese and Brazilian Portuguese during childhood is /i/. [14]. Yet, with our small sample size and in the absence of a language background questionnaire it is unwarranted to draw any conclusion. Future research should investigate whether the quality of the default epenthetic vowel can be influenced by early language exposure to Brazilian Portuguese even in reportedly monolingual Japanese speakers.

5. Acknowledgements

We wish to thank the undergraduate students at the Nagoya University of Commerce and Business who participated in our experiment. We also thank the three Australian English speakers who lent us their voices for recording. This research is supported by a research grant from the Japanese Society for the Promotion of Science (#20K13055).

6. References

[1] Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., and Mehler, J., "Epenthetic vowels in Japanese: A perceptual illusion?", Journal of Experimental Psychology: Human Perception and Performance, 25(6):1568, 1999.

[2] Dupoux, E., Parlato, E., Frota, S., Hirose, Y., and Peperkamp, S., "Where do illusory vowels come from?", Journal of Memory and Language, 64(3):199-210, 2011.

[3] Shaw, J. A. and Kawahara, S., "The lingual articulation of devoiced /u/ in Tokyo Japanese", Journal of Phonetics, 66:100-119, 2018.

[4] Kilpatrick, A., Kawahara, S., Bundgaard-Nielsen, R., Baker, B., and Fletcher, J., "Japanese perceptual epenthesis is modulated by transitional probability", Language and Speech, 64(1):203-223, 2021.

[5] Shannon, C. E., "A mathematical theory of communication", The Bell System Technical Journal, 27(3):379-423, 1948.

[6] Guevara-Rukoz, A., Lin, I., Morii, M., Minagawa, Y., Dupoux, E., and Peperkamp, S., "Which epenthetic vowel? Phonetic categories versus acoustic detail in perceptual vowel epenthesis", The Journal of the Acoustical Society of America, 142(2), 2017.

[7] Durvasula, K. and Kahng, J., "Illusory vowels in perceptual epenthesis: The role of phonological alternations", Phonology, 32(3):385-416, 2015.

[8] Hallé, P., Segui, J., Dominguez, A., Cuetos, F., Jaichenco, V., and Sevilla, Y., "Special is especial but stuto is not astuto: Perception of prothetic /e/ in speech and print by speakers of Spanish", in Psicolinguística en Español. Homenaje a Juan Seguí, 31-47, 2014.

[9] Davidson, L. and Shaw, J. A., "Sources of illusion in consonant cluster perception", Journal of Phonetics, 40(2):234-248, 2012.

[10] R Core Team, R., "R: A language and environment for statistical computing", Build 548. Computer software, 2024.

[11] Bates, D., Maechler, M., Bolker, B., and Walker, S., "Fitting linear mixed-effects models using lme4", Journal of Statistical Software, 67:1-48, 2015.

[12] [Fox, J. and Weisberg, S., An R Companion to Applied Regression (3rd ed.), Sage, 2019.

[13] Lenth, R., "Least-squares means: The R package lsmeans", Journal of Statistical Software, 69:1-33, 2016.

[14] Parlato-Oliveira, E., Christophe, A., Hirose, Y., and Dupoux, E., "Plasticity of illusory vowel perception in Brazilian-Japanese bilinguals", Journal of the Acoustical Society of America, 127(6):3738-3748, 2010.

Decoding Surprisal and Iconicity in American English

Alexander J. Kilpatrick[1], *Rikke L. Bundgaard-Nielsen*[2]

[1]Nagoya University of Commerce and Business, [2]University of Melbourne

alexander_kilpatrick@nucba.ac.jp, rikkieb@unimelb.edu.au

Abstract

This meta-study investigates how phonemic bigram surprisal and iconicity affect word processing in American English. It shows that high surprisal words are harder to process than words with lower levels of surprisal, and that iconic words are easier to process than arbitrary words. The results also show that both high surprisal and iconic words are associated with improved memory recall. Additionally, the study shows that longer words generally convey less information than shorter words, especially at word boundaries, and that iconic words are more likely to maintain their high surprisal irrespective of length. These findings suggest that language evolves to the cognitive processing limitations of speakers with bigram surprisal and iconicity playing important roles in this process.

Index Terms: speech processing; surprisal; iconicity; memory recall; age of acquisition

1. Introduction

Most psycholinguists assume that language processing efficiency reflects a balance between and individual's cognitive capacity and the complexity and quantity of the information that is received [e.g., 1]. This assumption has been extensively explored and theorized in research focusing on the word level [2, 3, 4, 5] and highlights the differential effects of *predictability* on language processing. C*ognitive load theory* [6], for instance, posits that cognitive capacity is finite and that unpredictable words are processed less efficiently than predictable ones [7, 8] because of requiring increased cognitive effort. Other research indicates that decreased predictability can facilitate lexical recall. For instance, the *lossy-context surprisal model* [9] suggests that while unpredictable words pose processing challenges due to misalignment with prior expectations, the additional cognitive effort enhances the memorability of these words. This suggests that recall and comprehension are affected differentially by predictability in the input.

Here, we test the effect of two variables that influence processing: iconicity—or form-to-meaning mapping—and surprisal—a measure of the quantity of information presented. Iconicity refers to the resemblance between a word's form and its meaning and has been demonstrated to play a crucial role in language development, cognition, and processing efficiency [10, 11, 12]. For instance, longitudinal studies indicate that both children and parents tend to use more iconic words during early childhood, with a transition towards using more arbitrary language occurring as cognitive development progresses [10]. Iconic words are also processed faster and more effortlessly than arbitrary words and recalled more reliably [13]. Moreover, iconicity facilitates word learning by establishing a connection between form and meaning based on resemblance, simplifying the learning process [14]. This facilitatory role of iconicity underscores its significance in language processing and acquisition.

Research has also explored the effect of *phonological iconicity* across various sensory domains in human language [15] including size [16], shape [17], and colour [18]. Iconic words often exhibit marked phonological traits such as the use of rare or foreign speech sounds (e.g., blech [blɛx]), phonotactic violations (e.g., vroom [vɹuːm]), expressive gemination (e.g., KAP-POW! [kəˈpːaʊ]), vowel lengthening (e.g., WHAAT? [wæːt]), and expressive metathesis (e.g., aks [æks] from ask) [19]. Here, the meta-data analysed is cross-referenced with a dictionary, so these examples do not feature in the dataset; however, we explore the idea that increased average surprisal is another way that iconic words exhibit markedness.

Despite considerable variation in the average *speaking rate* (phonemes over time) across the languages of the world, speakers of different languages transmit very similar volumes of *information* per unit of time. This optimization for efficient communication supports the cognitive economy principle discussed in the current study, where we posit that languages evolve to balance communicative efficiency with processing limitations. For example, a comparison between English and Japanese indicates that English speakers typically produce 6.19 syllables/second while Japanese speakers manage 7.84 syllables/second [20]. Conversely, English has a larger phonemic inventory than Japanese and allows greater phonotactic variation, which has the consequence that the average possibility of any two phonemes co-occurring being much lower in English than in Japanese, and that English expresses more information *per phoneme* than Japanese does. The slower rate of speech in English, however, results in *a similar overall rate of information transmission* between the two languages. The similar rate of information transmission between Japanese and English is not unique to this pairing. Indeed, [21] examined the relationship between speech rate and information expression in 17 languages and found that they trend towards encoding similar information rates (39 bits/s), and [21] proposes that this is evidence that information transmission rate is modulated by universal processing limitations. Similarly, [22] showed that less-probable words tend to have segments that provide more information early, facilitating quicker and more accurate identification in a cross-linguistic study.

The present study moves away from a focus on words in some prior research and examines how predictability at the *phoneme level* affects language comprehension in American English by examining their transitional probability. To do so we calculate average bigram surprisal (hereafter: average surprisal) which is Shannon's surprisal [23] calculated on the predictability of each phoneme given its prior. This returns values in bits of information where high information reflects low predictability. Average surprisal is calculated as the sum of information divided by the number of bigrams. This calculation is included in a series of models designed to measure the

ISSN 2207-1296 © 2024 ASSTA

influence of average surprisal on the results of pre-existing psycholinguistic tests. The results are consistent with psycholinguistic theories that suggest a trade-off between the cognitive load imposed by unexpected input and the facilitative effect of iconicity on language processing [1, 24]. In addition, we observe that while increased word length correlates with decreased surprisal at word boundaries, iconic words defy this trend and maintain high surprisal irrespective of length.

2. Method

2.1. The Master Dataset

The Master Dataset in the present study is comprised of the SUBLEX-US corpus [25] which was cross-referenced with the Carnegie Mellon University Pronouncing Dictionary [26] to convert English orthography to phonemic transcriptions based upon Standard American English pronunciation. Word length is the number of phonemes in each word. After surprisal was calculated, the master dataset was cross-referenced with additional datasets to obtain parts of speech [27], morpheme counts [28], and iconicity ratings [29] from a study in which 1400 American English speakers rated how similar each word "sounds like" its meaning on a 7-point Likert scale. In the master dataset, any word that did not find a match was discarded, resulting in a dataset of 39,136,598 instances of 13,336 unique words.

2.2. Psycholinguistic Datasets

We include datasets from five different published psycholinguistic studies: The first dataset consists of an auditory lexical decision test (MALD: [30]) completed by 231 American English-speaking participants. This study provides a dataset of reaction times and accuracy of recognizing real words, with 10,340 samples matching the master dataset, discussed above. The second dataset comes from a speeded reading experiment [31], with 816 native English-speaking participants recruited from six universities in the United States. We also include two datasets examining the age-of-acquisition of words: the first [32] involved 1960 American English participants indicating the age at which they believed they would have understood a given word, yielding 12,465 words that match the Master Dataset, while the second dataset [33] comprised responses from 829 participants at the University of Glasgow, rating words based on when they learned them, with 4101 matching words in the master dataset. Finally, we include data from a word recognition accuracy study [34] completed by 120 undergraduate students trained on a list of words in one experimental session and a subsequently tested on the accuracy of their recall in a second session within the same week.

2.3. Predictions

Hypothesis 1: We expect that word length is associated with decreased average surprisal. **Hypothesis 2:** We also expect to find that iconic words carry more information and that the effect of word length on average surprisal to be dampened in iconic words due to the inherent effortlessness of their processing. **Hypothesis 3:** We also expect to find some influence of word position on surprisal whereby the correlation between length and surprisal is dependent on the position of the bigram. **Hypothesis 4:** In reference to the psycholinguistic battery, we expect that iconic words will be processed more accurately and faster and learned at a younger age while high average surprisal words will be processed less accurately, more slowly, and learned at an older age. **Hypothesis 5:** Finally, we predict that both increased iconicity and average surprisal will be associated with increased memory recall.

3. Results

All models were constructed in R [35]: https://osf.io/mduc9/?view_only=e20f9f4d92d54044b0ab941122b3d24a. A linear regression model was constructed to examine the relationship between length and average surprisal (**H1**). The regression equation was significant, $F(1, 13704) = 501.5$, $p < .001$, with an R-squared value of .035, indicating that approximately 3.5% of the variance in average surprisal can be explained by length. The coefficient for average surprisal was significant, $b = -0.42$, $t(13704) = -22.39$, $p < .001$, indicating that average surprisal decreases as a function of length. A multiple regression analysis was conducted to predict average surprisal based on length, iconicity, number of morphemes, and parts of speech. The results are presented in Table 1.

Table 1. *Multiple linear regression model results. Asterisks denote statistical significance (* p < 0.05, ** p < 0.01, *** p < 0.001).*

Variable	Result
Intercept	33.456***
F statistic	66
Degrees Freedom	15:13288
Adjusted R2	0.069
Phonemic Length	-0.085
Iconicity	9.600***
Phonemic Length:Iconicity	-6.567***
Morphemic Length	9.527***
PoS_Adverb	-3.625***
PoS_Article	-3.943***
PoS_Conjunction	-1.679
PoS_Determiner	-3.035**
PoS_Interjection	5.136***
PoS_Name	-0.042
PoS_Noun	5.818***
PoS_Number	-2.858**
PoS_Prepositon	-3.675***
PoS_Pronoun	-5.542***
PoS_Verb	-3.203**

The regression equation was significant, $F(15, 13288) = 66.01$, $p < .001$, with an R-squared value of .069, indicating that approximately 6.9% of the variance in average surprisal can be explained by the predictors. Among the predictors, iconicity ($b = 0.26$, $p < .001$), number of morphemes ($b = 0.15$, $p < .001$), and the interaction between length and iconicity were significant predictors of average surprisal (**H2**). However, length alone was not a significant predictor of average surprisal ($b = -0.00$, $p = .932$). This suggests that iconic words carry more information than arbitrary words, additional morphemes are associated with increased average surprisal when phoneme count is controlled, and the effect of length in the previous model is explained entirely by the interaction between length and iconicity whereby average surprisal does not appear to be affected by increased length in iconic words.

To test **H3**, we generated heatmaps using bigram surprisal, phonemic length, and word position to visualize how information is expressed across words. Figure 1 reveals that increased length appears to only influence surprisal at word boundaries. We considered that this may be the result of affixes which are highly predictable sequences of sounds that occur at word boundaries in English. To test this possibility, we produced a second heatmap (Figure 2) using only monomorphemic words to control for affixation. The word-boundary pattern is consistent across heatmaps. We explored this further by running a series of simple linear regression analyses to assess the relationship between word length and bigram surprisal at word start, middle, and end positions. Longer words have lower surprisal at their onset ($\beta = -0.139$, $p < .001$, $R^2 = 0.033$), and similarly at their end ($\beta = -0.171$, $p < .001$, $R^2 = 0.043$), highlighting the significance of word boundaries in information expression. The Word Middle model showed a much smaller effect size ($\beta = -0.017$, $p = .027$, $R^2 < 0.001$), suggesting a weaker influence of word length on surprisal in the middle of words. These findings are consistent in monomorphemic words only.

To test **H4** and **H5**, a series of multiple linear regression models were constructed using the accuracy and reaction times of the auditory lexical decision task, the accuracy and reaction times of the speeded reading experiment, the age of acquisition (AoA) scores of the two AoA experiments, and the accuracy scores of the memory recall experiment (Table 2). The analyses are consistent with H4, demonstrating both that iconic words are processed more accurately and rapidly, and acquired at a younger age compared to non-iconic words, and that high surprisal words, are processed with lower accuracy, slower response times, and tend to be learned at an older age. Finally, **H5** is also supported, as both increased iconicity and average surprisal are associated with enhanced recall, highlighting the mnemonic advantage conferred by these linguistic features.

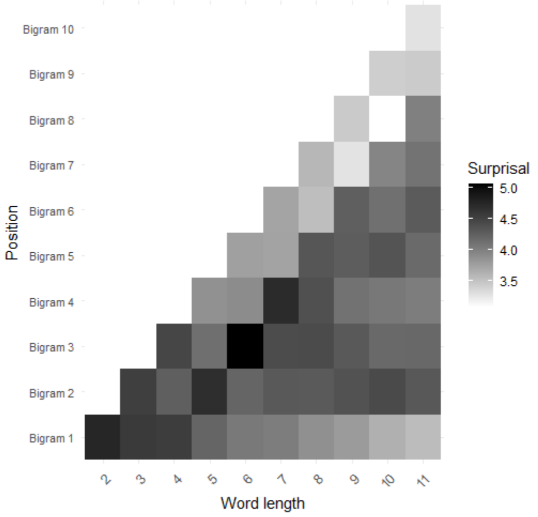


Figure 1: *Heatmap of surprisal across words according to length.*

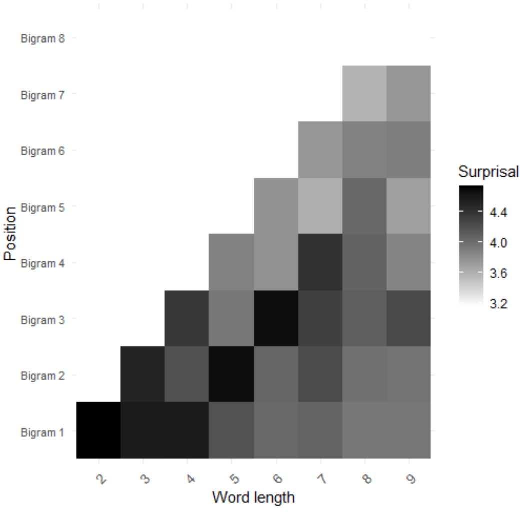


Figure 2: *Heatmap of the distribution of information according to length in monomorphemic words.*

Table 2. *Multiple linear regression model results constructed to test how various variables influence Average Surprisal. Asterisks denote statistical significance (* p < 0.05, ** p < 0.01, *** p < 0.001).*

Test	Lexical Decision		Reading Task		Age of Acquisition		Memory
Variable	Accuracy	RT	Accuracy	RT	AOA:[32]	AOA:[33]	Accuracy
Intercept	88.98***	67.69***	244.53***	122.09***	44.76***	20.84***	30.25***
F statistic	11.05	28.4	44.79	417.6	261.8	70.83	58.71
DF	15:10324	15:10324	15:13093	15:13093	15:12499	14:4086	11:4561
Adjusted R2	0.016	0.04	0.049	0.3228	0.239	0.195	0.122
Average Surprisal	**-3.02****	**3.05****	**-6.10*****	**9.37*****	**8.79*****	**6.62*****	**30.25*****
Iconicity	**3.1****	**-2.17***	**16.07*****	**-11.77*****	**-20.26*****	**-6.89*****	**8.09*****
Phonemic Length	6.46***	13.25***	-13.40***	56.65***	33.40***	21.38***	3.32***
Morphemic Length	1.97*	2.169*	5.81***	-4.44***	1.42	0.17	-2.36*
PoS_Adverb	1.33	-0.6	3.30***	-5.64***	-11.47***	-3.60***	-3.87***
PoS_Article	-4.64***	1.33	0.92	0.10	-2.57*	-1.54	
PoS_Conjunction	-1.39	1.02	0.27	-1.06	-5.09***		
PoS_Determiner	-0.36	0.42	1.98*	-1.56	-5.95***	-2.35*	
PoS_Interjection	-0.35	2.80**	-1.17	0.33	-3.81***	-2.54*	0.28
PoS_Name	0.68	2.49*	1.08	-0.32	-0.80	-1.01	2.57*
PoS_Noun	2.63**	0.73	5.42***	-3.53***	-8.40***	-2.24*	5.47***
PoS_Number	0.48	0.82	2.03*	-1.56	-8.48***	-2.44*	0.49
PoS_Prepositon	-1.6	0.77	1.92	-1.56	-6.53***	-2.71**	-1.06
PoS_Pronoun	1.22	0.016	2.56*	-3.93***	-10.06***	-1.97*	
PoS_Verb	2.13*	2.94**	4.84***	-0.60	1.16	0.37	-10.39***

4. Discussion

In the present investigation, we demonstrated that word length significantly influences surprisal within words, with longer words typically conveying less information per phoneme (**H1**). Interestingly, the effect of length on surprisal is largely explained by its interaction with iconicity (**H2**). This supports previous studies that suggest a unique processing advantage for iconic words [24]. Additionally, bigram position in words modulates information distribution (**H3**), with length's impact on surprisal particularly pronounced at word boundaries, possibly to facilitate early identification [22]; however, this does not explain why the effect was observed at the end of words and morphemes. Our analysis of the data from the psycholinguistic tests presented in [30, 31, 32, 33, 34] also demonstrate that highly iconic words are associated with increased processing efficiency and decreased age of acquisition, while high surprisal words exhibit the opposite pattern (**H4**). Both increased surprisal and iconicity correlate with improved recognition memory (**H5**), highlighting a complex relationship between linguistic properties and cognitive performance. These findings provide evidence for a complex relationship between cognitive cost and language efficiency, and indicate that cognitive constraints might be an important factor in language evolution. They also suggest that existing models, such as *cognitive load theory* [6] and the *lossy-context surprisal model* [9] should be extended beyond the word-level to include the transitional probabilities of phonemes. Within these existing frameworks, words consisting of predictable phoneme sequences are assumed to be easier to process while those made up of unpredictable sequences are more difficult. However, the additional investment of cognitive resources in these more difficult words enhances long-term recall. Iconic words are an important piece of this puzzle because they are processed with inherent effortlessness. Therefore, they can be used as a benchmark to measure the influence of increased cognitive costs by observing how little these effects impact iconic words, compared to their non-iconic counterparts. This includes the effect of word length on average surprisal which is entirely accounted for by its interaction with iconicity according to the model presented in Table 1.

The relationship between word length, iconicity, and average surprisal sheds light on how language is adapted to speakers' cognitive demands. The fact that longer words typically convey less information per phoneme can be argued to reflect a principle of linguistic efficiency. Despite this, iconic words, which have higher surprisal, are processed with greater accuracy and speed, suggesting a cognitive advantage for iconicity that is strong enough to counter the increased processing cost associated with higher levels of information. This preference for iconicity not only enhances cognitive processing and recall, but potentially also, influences how English has evolved. *The iconic treadmill hypothesis* [36], for example, posits that iconic words tend to evolve towards arbitrariness. Perhaps that process of evolution towards arbitrariness also creates more predictable sequences of sounds to offset the increased cognitive demands of processing non-iconic words. Consider the evolution of the English word *laugh*, which goes back to Old English *hlehhan* [37] where the iconic associations are much more evident. In old English, the hl-onset was very likely a low frequency/high surprisal sequence because it is no longer phonotactically legal. As the word evolved, it lost most of its iconic associations and became a comparatively predictable CVC sequence. This example speaks to how the transition from iconicity to arbitrariness in language may involve a reduction in information, making words easier to process.

Cognitive load is closely tied to predictability in language processing and *cognitive load theory* [6] suggests that the brain expends more cognitive resources to integrate unpredictable words into a given context than it does predictable words, resulting in longer reading times and slower response times. However, despite the increased processing effort, unpredictable words may be more memorable, as suggested by the *lossy-context surprisal model* [9]. This model proposes that the processing difficulty of a word is linked to its surprisal value within a memory representation of the context. High-surprisal words, although more challenging to process initially, may lead to deeper encoding and better memory recall due to the additional cognitive effort required for integration. Indeed, the experiments in our study support this idea, showing a positive correlation between memory recall and surprisal.

The investigation into how the position of bigrams within words influences surprisal reveals that the impact of word length on information is primarily observed at word boundaries. This would suggest that word boundaries may play a particularly important role in directing the allocation of cognitive resources by listeners for upcoming signals, if a word can be (correctly) retrieved from the lexicon/memory based on its initial segments only. However, this interpretation is preliminary and warrants further exploration to understand the significance of word boundaries as cues for cognitive resource allocation. This allows us to explore the interaction between linguistic structure and cognitive mechanisms. Additionally, it would be valuable to investigate potential differences between complex multi- and monomorphemic words. [38] highlights that complex words have multiple points at which the probability of a target word shifts. These points influence response latencies due to the Surprisal carried by phonemes suggesting that the cognitive cost of updating probability distribution differs between simple and complex words.

We recognize that our study, which focuses exclusively on English words, is of course limited. The results raise the question of whether similar patterns exist in other languages, with different phonemic inventories, different phonotactics, and differently shaped lexicons, prompting the need for further exploration. Future studies could broaden their scope to encompass languages from various linguistic families and cultural contexts. Such comparative analyses would not only deepen our comprehension of universal versus language-specific phenomena but also illuminate broader principles governing human language processing and evolution.

Our findings suggest that (at least) English has evolved in response to cognitive demands by minimizing cognitive load. Longer words tend to convey less information, particularly at word boundaries, but their impact diminishes with easier processing. Despite carrying more information, iconic words are processed more efficiently, illustrating a cognitive advantage. These insights, though limited to American English, prompt further investigation across diverse languages to better understand the universal versus language-specific effects on language processing and evolution.

5. Acknowledgements

We wish to thank Aleksandra Ćwiek, Brett Baker, Sharon Peperkamp, Shigeto Kawahara, David Sidhu, Maria Flaksman, and Kimi Akita for their insightful comments and suggestions for this project. We also thank the researchers who made their data publicly available so we could conduct this meta-analysis. This project was funded by the Japan Society for the Promotion of Science (# 20K13055).

6. References

[1] Bybee, J., “From usage to grammar: The mind's response to repetition,” Language, 82(4):711-733, 2006.

[2] Ehrlich, S. F., & Rayner, K., “Contextual effects on word perception and eye movements during reading,” Journal of Verbal Learning and Verbal Behavior, 20(6):641-655, 1981.

[3] Altmann, G. T., & Kamide, Y., “Incremental interpretation at verbs: Restricting the domain of subsequent reference,” Cognition, 73(3):247-264, 1999.

[4] Trueswell, J. C., Tanenhaus, M. K., & Kello, C., “Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths,” Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(4):800-819, 1994.

[5] Elman, J. L., Hare, M., & McRae, K., “Learning and the structure of syntactic categories,” in K. Johnson & E. N. Gibson [Eds.], Perceptual and cognitive aspects of language and speech, pp. 103-128, Springer, 2005.

[6] Sweller, J., “Cognitive load during problem solving: Effects on learning,” Cognitive Science, 12(2):257-285, 1988.

[7] Federmeier, K. D., & Kutas, M., “A rose by any other name: Long-term memory structure and sentence processing,” Journal of Memory and Language, 41(4):469-495, 1999.

[8] Kutas, M., & Federmeier, K. D., “Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP),” Annual Review of Psychology, 62:621-647, 2011.

[9] Futrell, R., Gibson, E., & Levy, R. P., “ Lossy ‐ context surprisal: An information ‐ theoretic model of memory effects in sentence processing,” Cognitive Science, 44(3), 2020.

[10] [Perlman, M., Fusaroli, R., Fein, D., & Naigles, L., “The use of iconic words in early child-parent interactions,” in the 39th Annual Conference of the Cognitive Science Society (CogSci 2017), pp. 913-918, Cognitive Science Society, 2017.

[11] Edmiston, P., Perlman, M., & Lupyan, G., “Repeated imitation makes human vocalizations more word-like,” Proceedings of the Royal Society B: Biological Sciences, 285(1874):20172709, 2018.

[12] Nölle, J., Fusaroli, R., & Tylén, K., “Iconicity in sign grounding: Representation or disambiguation,” in The Evolution of Language: Proc. of the 13th Int. Conf. Evolution of Language (EVOLANG XIII), pp. 318-320, 2020.

[13] Sidhu, D. M., Khachatoorian, N., & Vigliocco, G., “Effects of Iconicity in Recognition Memory,” Cognitive Science, 47(11), 2023.

[14] Nielsen, A. K., & Dingemanse, M., “Iconicity in word learning and beyond: A critical review,” Language and Speech, 64(1):52-72, 2021.

[15] Perniss, P., & Vigliocco, G., “Iconicity: A review,” in The Oxford Handbook of Synesthesia, pp. 947-973, Oxford University Press, 2014.

[16] Sapir, E., “A study in phonetic symbolism,” Journal of Experimental Psychology, 12(3):225-239, 1929.

[17] Ćwiek, A., et al., “The bouba/kiki effect is robust across cultures and writing systems,” Philosophical Transactions of the Royal Society B, 377(1841):20200390, 2022.

[18] Erben Johansson, N., The building blocks of sound symbolism, Doctoral dissertation, Lund University, 2020.

[19] Voeltz, F. K. E., & Kilian-Hatz, C. [Eds.], Ideophones: Typological studies in language, Vol. 44, John Benjamins, 2001.

[20] Pellegrino, F., Coupé, C., & Marsico, E., “Across-language perspective on speech information rate,” Language, 87(3):539-558, 2011.

[21] Coupé, C., et al., “Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche,” Science Advances, 5(9), 2019.

[22] King, A., & Wedel, A., “Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing,” Open Mind, 4:1-12, 2020.

[23] Shannon, C. E., “A mathematical theory of communication,” Bell System Technical Journal, 27(3):379-423, 1948.

[24] Sidhu, D. M., Vigliocco, G., & Pexman, P. M., “Iconicity in language processing and learning,” Annual Review of Psychology, 71:633-663, 2020.

[25] Brysbaert, M., & New, B., “Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English,” Behavior Research Methods, 41(4):977-990, 2009.

[26] Weide, R., “The Carnegie Mellon pronouncing dictionary,” Release 0.6, 1998.

[27] Brysbaert, M., New, B., & Keuleers, E., “Adding part-of-speech information to the SUBTLEX-US word frequencies,” Behavior Research Methods, 44:991-997, 2012.

[28] Sánchez-Gutiérrez, C. H., et al., “MorphoLex: A derivational morphological database for 70,000 English words,” Behavior Research Methods, 50:1568-1580, 2018.

[29] Winter, B., et al., “Iconicity ratings for 14,000+ English words,” Behavior Research Methods, 1-16, 2023.

[30] Tucker, B. V., et al., “The massive auditory lexical decision (MALD) database,” Behavior Research Methods, 51:1187-1204, 2019.

[31] Balota, D. A., et al., “The English lexicon project,” Behavior Research Methods, 39:445-459, 2007.

[32] Kuperman, V., et al., “Age-of-acquisition ratings for 30,000 English words,” Behavior Research Methods, 44:978-990, 2012.

[33] Scott, G. G., et al., “The Glasgow Norms: Ratings of 5,500 words on nine scales,” Behavior Research Methods, 51:1258-1270, 2019.

[34] Cortese, M. J., et al., “Recognition memory for 2,578 monosyllabic words,” Memory, 18(6):595-609, 2010.

[35] R Core Team, “R: A language and environment for statistical computing,” Build 548, 2024.

[36] Flaksman, M., “Iconic treadmill hypothesis: The reasons behind continuous onomatopoeic coinage,” in A. Zirker, et al. [Eds.], Dimensions of Iconicity, pp. 15-38, John Benjamins, 2017.

[37] Wedgwood, H., & Atkinson, J.C., A dictionary of English etymology, Trübner & Company, 1872.

[38] Balling, L. W., & Baayen, R. H., “Probability and surprisal in auditory comprehension of morphologically complex words,” Cognition, 125(1):80-106, 2012.

Acoustic Analysis of Vowel Production Using Magnetic Resonance Imaging

Tharinda Piyadasa[1], Michael Proctor[2], Amelia Gully[3], Yaoyao Yue[1], Kirrie Ballard[4], Naeim Sanaei[5], Sheryl Foster[4, 5], Tünde Szalay[2], David Waddington[4], Craig Jin[1]

[1]School of Electrical and Computer Engineering, University of Sydney, Australia
[2]Department of Linguistics, Macquarie University, Australia
[3]Department of Language and Linguistic Science, University of York, UK
[4]Faculty of Medicine and Health, University of Sydney, Australia
[5]Radiology Department, Westmead Hospital, Australia

tharinda.piyadasa@sydney.edu.au

Abstract

Details of individual speaker vocal tract configurations remain understudied due to the limitations of most instrumental phonetic methods. Midsagittal articulation of /iː-ɑː-ɔː-uː-ɜː/ by a speaker of Southern Standard British English was captured using real-time Magnetic Resonance Imaging. Three-dimensional tract configurations during production of the same vowels were acquired using high-resolution volumetric imaging. Acoustic models derived from imaging data were compared with reference acoustic recordings. Models demonstrate particular sensitivity to palatal and velar tract geometry; details of pharyngeal structures had less influence on acoustic responses. These data demonstrate the importance of multimodal data in acoustic characterization of individual speaker vowels.

Index Terms: vowels, MRI, English, vocal tract, area function, acoustic modeling

1. Introduction

Understanding how the vocal tract is configured during vowel production has been a central concern of speech science and a foundational topic informing models of speech production [1, 2, 3, 4]. Magnetic resonance imaging (MRI) has advanced the study of vowel production by allowing safe, flexible and accurate imaging of soft tissue [5, 6, 7]. Vocal tract geometries have been resolved in detail by orienting image planes perpendicular to the axis of airway [8, 9, 10, 11], and volumetric imaging techniques have provided comprehensive coverage of the whole upper airway at increasing spatial resolutions [12, 13, 14]. These methods have revealed the complex geometries involved in vowel production and how they vary between speakers, languages, and allophones [15, 16, 17], informing more detailed vocal tract representations beyond idealized tube models [18, 19, 20, 21, 22]. These data are advancing our understanding of the complex relationships between vocal tract morphology, articulation, and acoustics [23, 24, 25], but many aspects of vocal tract shaping and its acoustic consequences are still imperfectly understood.

High resolution volumetric imaging of the vocal tract can be achieved using multi-second acquisition times, but because participants must sustain vowels in these studies, the resulting postures are static, and may be hyperarticulated. Real-time MRI (rtMRI) allows imaging of the upper airway during connected speech produced with more natural prosody [26, 27], which is important for phonetic characterization of vowels [28, 29, 30]. rtMRI has provided insights into the dynamics of vowel articulation in French, Portuguese, English, and other languages [31, 32, 33, 34]. Note that for speech MRI studies, the supine participant posture may affect vocal tract shape [35], and the loud noise means that Lombard speech is usually captured [36].

Volumetric and real-time imaging offer important complementary insights into vowel production, but reconciling data from different modalities creates additional challenges. Companion volumetric and real-time MRI data have previously been combined to examine vocal tract shaping in Swedish and French vowels [31, 16]. Articulatory data obtained using different sensing methods can be assessed by comparing acoustic responses of models derived from the corresponding vocal tract configurations. Acoustic responses have been estimated from MRI data of Czech, Finish and English vowels using Finite Element [37, 38, 39, 40], finite-difference time-domain [20], and 3D digital waveguide methods [41, 42]. While these techniques rely on multi-dimensional representations of the vocal tract, acoustic responses can be estimated from 1D tract models [43, 44], allowing for direct comparison of models based on articulatory data captured during vowel production with acoustic recordings of the corresponding vowels produced by the same speaker.

The goal of this study is to examine details of vowel production in a speaker of British English in new detail using volumetric MRI, real-time MRI, and acoustic recordings. We explore the acoustic responses of vocal tract configurations by synthesizing vowels from area functions derived from MRI data, and validate these acoustic models against out-of-scanner reference recordings of vowels produced by the same speaker. By comparing acoustic outputs of vocal tract models derived from each dataset, we aim to assess the relative advantages of each imaging modality for acoustic modelling of vowels, and methods for extracting vocal tract representations appropriate for vowel models from each type of data. Finally, we assess the impact on acoustic modelling of different methods of representation of data obtained from each imaging modality, and examine how different approaches to segmentation of vocal tract boundaries affect acoustic responses of tract models.

2. Methods

Data were collected during the pilot phase of a larger project examining development of speech motor control in adolescents. An adult female L1 speaker of Standard Southern British English produced vowels in a series of speech tasks recorded out of and inside an MRI scanner. Vowels /iː-ɑː-ɔː-uː-ɜː/ were elicited

Table of Contents for this manuscript

in monosyllabic words *"beet', 'Bart', 'bought', 'boot', 'Bert"*. Each token was recorded once in a quiet room with a Glottal Enterprises EG2-PCX2 digital speech recorder to familiarize the participant with the experimental materials. The same utterances were later recorded five times during a real-time MRI scan, and additionally as sustained productions during a volumetric MRI scan. A total of (5 words) × (1 pre-scan + 5 rtMRI + 1 volumetric MRI) = 35 vowel exemplars were included in the analysis.

2.1. Vocal Tract Imaging

MRI data were acquired at Westmead Hospital on a Siemens Magnetom Prisma 3T scanner with a 64-channel head/neck receiver array coil. The speaker's upper airway was imaged while lying supine. Data were acquired from an 8 mm slice aligned with the mid-sagittal plane, over a 280 × 280 mm field of view, using a 2D RF-spoiled, radially-encoded FLASH sequence [45]. Audio was recorded concurrently in-scanner at 16 kHz using an Opto-acoustics FOMRI-III ceramic noise-canceling microphone designed for MRI environments [46]. rtMRI data were reconstructed in Matlab into midsagittal videos with a pixel resolution of 0.97 mm^2, encoded as 72 frames per second MP4 files. Audio and video were time-aligned during postprocessing and video reconstruction.

3D configuration of the vocal tract during sustained (7.6 s) vowel production was captured using volumetric imaging of the upper airway. Data were acquired using a T1-weighted fast 3D gradient-echo sequence, with a spatial resolution of 160 × 160 × 32 px over a 256 × 256 × 64 mm field of view centred on the pharynx. These data provide detailed imaging of the entire upper airway, extending vertically from the upper trachea to the nasal cavities and sagittally from cheek to cheek, with a voxel resolution of 1.6 × 1.6 × 2.0 mm.

2.2. Vocal tract segmentation

Volumetric data (DICOM format), were processed using ITK-SNAP [47], an open-source tool for 3D segmentation of medical images. Contrast was enhanced, and the Snake tool was used for semi-automatic segmentation of vocal tract boundaries, from which 3D tract outlines were extracted (Fig. 1).

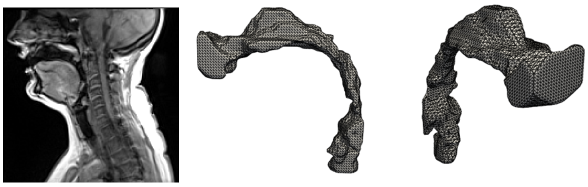


Figure 1: *Vocal tract configuration, sustained [ɑː].* Midsagittal slice and 3D volume extracted from segmented volumetric data

rtMRI data were analyzed using *inspect_rtMRI*, a Matlab-based graphical interface for inspection and semi-automatic segmentation of rtMRI data [48]. Midsagittal vocal tract boundaries were located in image frames corresponding to articulatory target postures for each vowel, and area functions were extracted at 7.7 mm intervals, from glottis to labial midpoint. Midsagittal slices were extracted from 3D vocal tract models, and an additional set of vocal tract area functions were calculated using the same method, to obtain a second set of high resolution midsagittal vocal tract representations for each vowel (Fig. 2).

2.3. Modelling Acoustic Responses

Each vocal tract (VT) was modeled as a series of concatenated cylindrical tubes with cross-sectional areas (CSA) derived from area functions, normalized by π. Tube models were downsampled through linear interpolation:

$$N = \text{round}\left(\frac{\text{len}(\texttt{VT}) \times \frac{Fs}{2} \times 4}{c}\right)$$

where *Fs* = 16 kHz, *c* = 350 m/s (speed of sound in moist air at body temperature 37°C [2]).

Reflection coefficients were calculated for each segment junction within the vocal tract model to simulate acoustic impedance mismatches [49]. The number of reflection coefficients corresponds to the length of the interpolated tube model, where the coefficients were derived using the formula:

$$r = \frac{A_{i+1} - A_i}{A_{i+1} + A_i}$$

where A_i and A_{i+1} are the CSAs of adjacent tube sections. Additionally, a finite lossy tube was modeled by appending a reflection coefficient to represent the mouth's impedance, with the value set to 0.71 [50]. A Rosenberg glottal pulse [51] was generated and processed through the vocal tract filter designed by converting reflection coefficients into filter coefficients using Durbin's recursion:

$$a_{k+1}[n] = a_k[n] + r_{k+1} \times a_k[k-n] \quad \text{for } n = 0, 1, \ldots, k$$

where $a_k[n]$ are the filter coefficients at recursion step k, r_{k+1} is the reflection coefficient at the $k+1$-th junction, and n indexes the coefficients in the filter.

The output speech signal was generated by convolving the glottal pulse with the acoustic filter coefficients, followed by amplitude normalization.

Acoustic properties of synthesized and recorded speech were compared using formant frequencies. F1, F2, F3 were tracked over speech intervals containing target vowels following the approach proposed in [52], using 20 ms Hamming analysis windows, 50% overlap, *max_F34cutoff* = 4500 Hz, and a pre-emphasis filter factor of 0.98.

3. Results and Discussion

Formant frequencies for vowels produced by the participant in reference (out-of-scanner) recordings were first compared to mean values (Table 1) reported for female speakers in Standard Southern British English [53]. Formants generally align closely with SSBE means; the participant's /iː/ is more fronted, and /ɔː/ is lower. Overall, formant values for short pronunciations are closer to SSBE means compared to sustained pronunciations, which may be attributed to the effects of hyperarticulation in sustained vowels. In particular, the large difference (30%) in F2 values for /uː/ shows that the sustained vowel was produced with a backed, more peripheral articulation.

Compared to out-of-scanner recordings, the in-scanner recordings typically exhibit larger F1 and F2 values ($\geq$ 4% difference). This is in line with the findings of [36], where it was established that increases in F1 and F2 occur due to scanner noise, and additional F1 increases can be attributed to the supine posture of the subject (Table 2). These effects were found to be subject-dependent. In this case, the scanner environment caused the tongue to be positioned higher and more

Table of Contents for this manuscript

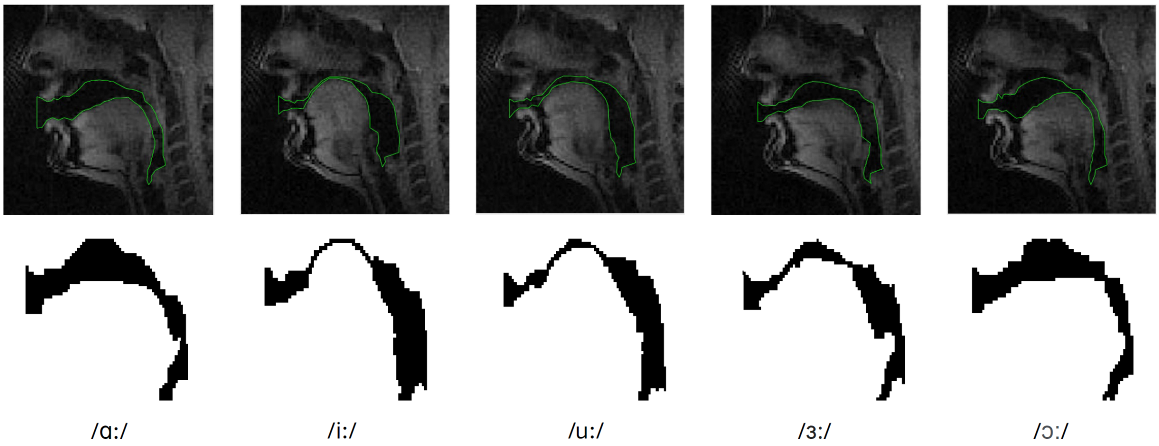


Figure 2: *Vocal tract segmentations used to calculate area functions*: Top: vowel target frames in rtMRI data; Bottom: midsagital sections from 3D tract volumes of sustained vowel postures. L-to-R: [ɑː-iː-uː-ɜː-ɔː]

Table 1. *Comparison of participant reference vowel formants with mean SSBE female vowel formant frequencies (Hz) [53]*

		F1	F2	F3
/ɑː/	SSBE Mean (F)	**910**	**1316**	**2841**
	Out-scanner (short)	-74	-152	+238
	Out-scanner (sustained)	-171	-40	-32
/iː/	SSBE Mean (F)	**303**	**2654**	**3203**
	Out-scanner (short)	+35	+144	-67
	Out-scanner (sustained)	-12	+202	+90
/uː/	SSBE Mean (F)	**328**	**1437**	**2674**
	Out-scanner (short)	+73	+60	+59
	Out-scanner (sustained)	+68	-436	+261
/ɜː/	SSBE Mean (F)	**606**	**1695**	**2839**
	Out-scanner (short)	-31	-112	+214
	Out-scanner (sustained)	-57	+36	+148
/ɔː/	SSBE Mean (F)	**389**	**888**	**2796**
	Out-scanner (short)	+18	-128	-796
	Out-scanner (sustained)	+110	-123	+354

Table 2. *Comparison of Formant Values Between In-Scanner Recordings and 1D Acoustic Model (rtMRI) with Out-Scanner Recordings (Short)*

		F1	F2	F3
/ɑː/	Out-scanner (short)	**836**	**1164**	**3079**
	In-scanner	+36	+72	-331
	1D acoustic model (rtMRI)	-80	+432	-214
/iː/	Out-scanner (short)	**338**	**2798**	**3136**
	In-scanner	+95	-492	-449
	1D acoustic model (rtMRI)	-44	-599	-546
/uː/	Out-scanner (short)	**401**	**1497**	**2733**
	In-scanner	+28	+243	-187
	1D acoustic model (rtMRI)	-37	+354	-53
/ɜː/	Out-scanner (short)	**575**	**1583**	**3053**
	In-scanner	+362	+58	-212
	1D acoustic model (rtMRI)	-72	+61	-234
/ɔː/	Out-scanner (short)	**407**	**760**	**2000**
	In-scanner	+208	+112	+576
	1D acoustic model (rtMRI)	+63	+424	+710

forward, reflecting a more constrained vocal tract shape during in-scanner recordings. In contrast, the 1D acoustic models based on rtMRI typically exhibit lower F1 values which may arise from the simplifications in the modeling process that fail to fully capture the open vocal tract configuration. Additionally, the acoustic model tends to have higher F3 values compared to in-scanner recordings, indicating differences in the back cavity configuration. This difference can be attributed to the back cavity segmentation being influenced by the presence of soft tissue, particularly around the epiglottis area.

Formants from the 1D acoustic models based on midsagittal volumetric images generally align with out-of-scanner sustained vowel recordings, though there are some notable discrepancies (Table 3). For instance, /uː/ and /ɔː/ show considerable differences in F2 values, with the 1D acoustic model having much higher values ($\geq$ 50% difference) compared to the corresponding out-of-scanner recordings. However, it should be noted that when compared to SSBE mean values and out-of-scanner values for short utterances, the out-of-scanner sustained values are much lower (Table 1). Also, the large difference in F3 values for /ɜː/ and /ɔː/ (16% and 22% difference respectively) suggest variations in the pharyngeal cavity shape, as from Figure 2

Overall, the F1 values for both 1D acoustic models are close to the out-of-scanner values, indicating a reasonable approximation of vertical tongue positions. However, F2 and F3 values exhibit greater deviations. Furthermore, the models often show smaller formant values compared to out-of-scanner recordings, which may be due to the lack of lip radiation effects in the synthesized speech.

3.1. Refinements in 3D midsagittal slices

Several adjustments were made to the 3D midsagittal segmentations to observe the accuracy of our acoustic modeling. These adjustments involved refining soft tissue boundaries in the regions around the hard palate, velar constrictions, and epiglottis. The changes were prompted by initial observations that revealed anatomical inaccuracies in the 3D segmentations, such as an unusually large cavity at the hard palate in /ɑː/ and /ɔː/, likely due to the relatively smaller amounts of soft tissue affecting the

Table 3. *Comparison of Formant Values Between 1D Acoustic Model (Volumetric) with Out-Scanner Recordings (Sustained)*

		F1	F2	F3
/ɑː/	Out-scanner (sustained)	**739**	**1276**	**2809**
	1D acoustic model (volumetric)	-65	+138	-187
/iː/	Out-scanner (sustained)	**291**	**2856**	**3293**
	1D acoustic model (volumetric)	+51	-630	-68
/uː/	Out-scanner (sustained)	**396**	**1001**	**2935**
	1D acoustic model (volumetric)	0	+825	+265
/ɜː/	Out-scanner (sustained)	**549**	**1731**	**2987**
	1D acoustic model (volumetric)	-97	-76	-476
/ɔː/	Out-scanner (sustained)	**499**	**765**	**3150**
	1D acoustic model (volumetric)	+54	+379	-708

Table 4. *Comparison of Formant Values for Original and Refined Midsagittal Slices of 3D Volumetric Representations of Vowels /ɑː/ /ɔː/, and /ɜː/*

		F1	F2	F3
/ɑː/	Out-scanner (sustained)	**739**	**1276**	**2809**
	1D acoustic model (Volumetric)	-65	+138	-187
	Refined hard palate	-111	+123	+22
	Non-refined epiglottis	-128	+176	-198
/ɔː/	Out-scanner (sustained)	**499**	**765**	**3150**
	1D acoustic model (Volumetric)	+54	+379	-708
	Refined hard palate	+28	+366	-921
	Increased constriction	+10	-29	-934
/ɜː/	Out-scanner (sustained)	**549**	**1731**	**2987**
	1D acoustic model (Volumetric)	-97	-76	-476
	Non-refined epiglottis	-134	-76	-685

resolution of the upper airway boundary. All adjustments were made through manual post-processing of the initial segmentations located using ITK-SNAP (Sec. 2.2).

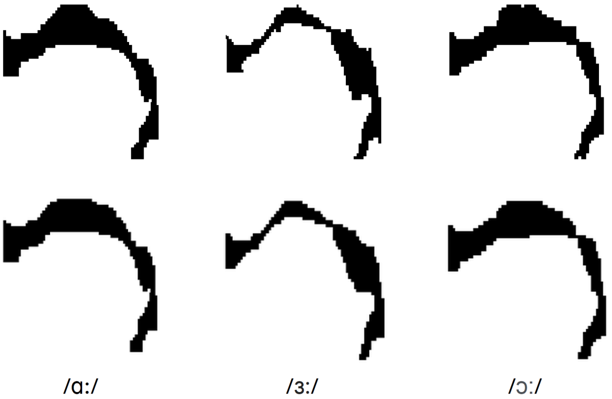


Figure 3: *Refinements in 3D midsagittal slices*: Top: original midsagittal sections; Bottom: midsagital sections after manual post-processing. L-to-R: [ɑː-ɜː-ɔː]

The adjustments made to the 3D midsagittal segmentations led to closer alignment of the formant values with the out-of-scanner recordings (Table 4). Refining the hard palate in /ɑː/ and /ɔː/ generally improved the alignment of F1 and F2 values by reducing exaggerated resonances caused by an initially larger hard palate. A similar improvement was observed when the velar constriction was increased in /ɔː/. This was expected, as increased velar constriction lengthens the front cavity of the vocal tract, thereby reducing F1 and F2 values.

The impact of the epiglottis definition in /ɑː/ and /ɜː/ was investigated to determine whether simplifying the epiglottis representation, as seen in rtMRI area functions, would improve the formant values (Table 4). However, this adjustment did not noticeably improve the formant values for either vowel. While the epiglottis influences the shape of the pharyngeal cavity, its impact on F1 and F2 is less pronounced compared to velar and palatal constrictions.

4. Conclusions

A method for determining detailed vocal tract configurations associated with vowel production has been proposed and validated using an acoustic synthesis framework. The impact of different tissue segmentation strategies has been assessed using 1D area functions extracted from rtMRI images and midsagittal slices of 3D data, validating these against both in-scanner and out-of-scanner acoustic recordings. The analysis showed that both tube models show variations in formant frequencies compared to out-of-scanner recordings. Overall, acoustic models based on midsagittal slices of 3D volumetric data more accurately represent natural speech formants compared to the models based on rtMRI images, indicating the importance of a better representation of the vocal tract geometry. The findings suggest that further improvements should include comprehensive vocal tract models considering lip radiation and dental structures. Therefore, future work will focus on using complete 3D vocal tract models to obtain acoustic responses and incorporating dental features and refined lip radiation models to improve vocal tract modeling.

5. Acknowledgements

Supported by Australian Research Council Discovery Grant DP220102933.

6. References

[1] T. Chiba and M. Kajiyama, *The vowel – its nature and structure*. Tokyo: Kaseikan, 1941.

[2] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Springer-Verlag, Berlin, 1972.

[3] G. Fant, *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. s'Gravenhage: Mouton, 1960.

[4] K. N. Stevens, *Acoustic Phonetics*. Cambridge: MIT Press, 2000.

[5] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *JASA*, vol. 90, no. 2, pp. 799–828, 1991.

[6] C. A. Moore, "The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images," *JSLHR*, vol. 35, no. 5, pp. 1009–1023, 1992.

[7] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *JASA*, vol. 100, no. 1, pp. 537–554, 1996.

[8] M. Matsamura, "Measurement of three-dimensional shapes of vocal tract and nasal cavity using magnetic resonance imaging," in *Proc. ICLSP*, 1992, pp. 779–782.

[9] D. Demolin, T. Metens, and A. Soquet, "Three-dimensional Measurement of the Vocal Tract by MRI," *ICLSP*, pp. 272–275, 1996.

[10] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images," *JPhon*, vol. 30, no. 3, pp. 533–553, 2002.

[11] P. Badin, G. Bailly, M. Raybaudi, and C. Segebarth, "A three-dimensional linear articulatory model based on MRI data," in *Proc. 3rd ETRW on Speech Synthesis*, 1998.

[12] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions for an adult female speaker based on volumetric imaging," *JASA*, vol. 104, no. 1, pp. 471–487, 1998.

[13] K. C. Welch, G. D. Foster, C. T. Ritter, T. A. Wadden, R. Arens, G. Maislin, and R. J. Schwab, "A novel volumetric magnetic resonance imaging paradigm to study upper airway anatomy," *Sleep*, vol. 25, no. 5, pp. 532–542, 2002.

[14] P. Badin and A. Serrurier, "Three-dimensional modeling of speech organs: Articulatory data and models," in *Tech. Comm. Psychological and Physiological Acoustics*, vol. 36, no. 5. Acoust. Soc. Japan, 2006, pp. 421–426.

[15] M. K. Tiede, "An MRI-based study of pharyngeal volume contrasts in Akan and English," *J. Phon.*, vol. 24, no. 4, pp. 399–421, 1996.

[16] O. Engwall, V. Delvaux, and T. Metens, "Interspeaker variation in the articulation of nasal vowels," *Proc. 7th ISSP*, pp. 3–10, 2006.

[17] Y. Wang, J. Dang, X. Chen, J. Wei, H. Wang, and K. Honda, "An MRI-based acoustic study of Mandarin vowels," in *Interspeech*, 2013, pp. 568–571.

[18] P. Mokhtari, T. Kitamura, H. Takemoto, and K. Honda, "Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients," *JPhon*, vol. 35, no. 1, pp. 20–39, 2007.

[19] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *JASA*, vol. 5, pp. 3231–3254, 2005.

[20] H. Takemoto, P. Mokhtari, and T. Kitamura, "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method," *JASA*, vol. 128, no. 6, pp. 3724–3738, 2010.

[21] S. Stone, M. Marxen, and P. Birkholz, "Construction and evaluation of a parametric one-dimensional vocal tract model," *IEEE/ACM Trans. ASLP*, vol. 26, no. 8, pp. 1381–1392, 2018.

[22] P. Birkholz, S. Kürbis, S. Stone, P. Häsner, R. Blandin, and M. Fleischer, "Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties," *Scientific data*, vol. 7, no. 1, p. 255, 2020.

[23] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoustical science and technology*, vol. 26, no. 1, pp. 16–26, 2005.

[24] A. Lammert, M. I. Proctor, A. Katsamanis, and S. S. Narayanan, "Morphological variation in the adult vocal tract: a modeling study of its potential acoustic impact," in *Interspeech*, Florence, Italy, 27-31 Aug. 2011, pp. 2813–2816.

[25] A. J. Gully, "Quantifying vocal tract shape variation and its acoustic impact: A geometric morphometric approach." in *Interspeech*, 2021, pp. 3999–4003.

[26] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *JASA*, vol. 115, no. 4, pp. 1771–1776, 2004.

[27] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K. S. Nayak, and S. Narayanan, "Analysis of speech production real-time MRI," *Comput Speech Lang*, vol. 52, pp. 1 – 22, 2018.

[28] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *JASA*, vol. 97, no. 5, pp. 3099–3111, 1995.

[29] M. Hashi, J. R. Westbury, and K. Honda, "Vowel posture normalization," *JASA*, vol. 104, no. 4, pp. 2426–2437, 1998.

[30] G. Morrison and P. Assmann, *Vowel Inherent Spectral Change*. Berlin: Springer, 2012.

[31] D. Demolin, S. Hassid, T. Metens, and A. Soquet, "Real-time MRI and articulatory coordination in speech," *Comptes Rendus Biologies*, vol. 325, no. 4, pp. 547–556, 2002.

[32] A. Teixeira, P. Martins, C. Oliveira, C. Ferreira, A. Silva, and R. Shosted, "Real-Time MRI for Portuguese," in *Computational Processing of the Portuguese Language*, H. Caseli, Ed. Berlin: Springer, 2012, pp. 306–317.

[33] M. Proctor, C. Lo, and S. Narayanan, "Articulation of English Vowels in Running Speech: a Real-time MRI Study," in *Proc. ICPhS*, Glasgow, 10-14 Aug. 2015.

[34] C. Carignan, R. K. Shosted, M. Fu, Z.-P. Liang, and B. P. Sutton, "A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French," *JPhon*, vol. 50, pp. 34–51, 2015.

[35] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech MRI: morphology and function," *Physica Medica*, vol. 30, no. 6, pp. 604–618, 2014.

[36] A. J. Gully, P. Foulkes, P. French, P. Harrison, and V. Hughes, "The Lombard effect in MRI noise," in *Proc. ICPhS*, 2019, pp. 5–9.

[37] P. Kršek, "Design of FE models of vocal tract for Czech vowels," in *Proc. Interaction and Feedbacks*, 2000, pp. 103–110.

[38] K. Dedouch, J. Horácek, T. Vampola, J. Švec, P. Kršek, and R. Havlık, "Acoustic modal analysis of male vocal tract for Czech vowels," *Interaction and Feedbacks*, pp. 13–19, 2002.

[39] D. Aalto, O. Aaltonen, R.-P. Happonen, P. Jääsaari, A. Kivelä, J. Kuortti, J.-M. Luukinen, J. Malinen, T. Murtola, R. Parkkola *et al.*, "Large scale data acquisition of simultaneous mri and speech," *Applied Acoustics*, vol. 83, pp. 64–75, 2014.

[40] M. Arnela, R. Blandin, S. Dabbaghchian, O. Guasch, F. Alías, X. Pelorson, A. Van Hirtum, and O. Engwall, "Influence of lips on the production of vowels based on finite element simulations and experiments," *JASA*, vol. 139, no. 5, pp. 2852–2859, 2016.

[41] M. Speed, D. Murphy, and D. Howard, "Modeling the vocal tract transfer function using a 3d digital waveguide mesh," *IEEE/ACM Trans. ASLP*, vol. 22, no. 2, pp. 453–464, 2014.

[42] A. J. Gully, H. Daffern, and D. T. Murphy, "Diphthong synthesis using the dynamic 3D digital waveguide mesh," *IEEE/ACM Trans. ASLP*, vol. 26, no. 2, pp. 243–255, 2018.

[43] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *The Bell System Technical Journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[44] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Comm.*, vol. 1, no. 3-4, pp. 199–229, 1982.

[45] A. J. Kennerley, D. A. Mitchell, A. Sebald, and I. Watson, "Real-time magnetic resonance imaging: mechanics of oral and facial function," *Br J Oral Max Surg*, vol. 60, no. 5, pp. 596–603, 2022.

[46] Optoacoustics Ltd., "FOMRI-II version 2.2," 2007.

[47] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[48] M. I. Proctor, D. Bone, and S. S. Narayanan, "Rapid semi-automatic segmentation of real-time Magnetic Resonance Images for parametric vocal tract analysis," in *Interspeech*, Makuhari, 26-30 Sept. 2010, pp. 1576–1579.

[49] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, A. V. Oppenheim, Ed. Prentice-Hall, 1978.

[50] A. Beköz, "Modeling of plosive to vowel transitions," Master's thesis, Middle East Technical University, 2007.

[51] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *JASA*, vol. 49, no. 2B, pp. 583–590, 1971.

[52] T. M. Nearey, P. F. Assmann, and J. M. Hillenbrand, "Evaluation of a strategy for automatic formant tracking," *JASA*, vol. 112, no. 5_Supplement, pp. 2323–2323, 2002.

[53] D. Deterding, "The formants of monophthong vowels in Standard Southern British English pronunciation," *JIPA*, vol. 27, no. 1-2, pp. 47–55, 1997.

Static and Dynamic Analyses of the Back Vowels /uː/ and /ʊ/ in Singapore English

Canaan Zengyu Lan, Olga Maxwell, Chloé Diskin-Holdaway

The University of Melbourne

canaan.lan@unimelb.edu.au;omaxwell@unimelb.edu.au;chloe.diskinholdaway@unimelb.edu.au

Abstract

This study explores static and dynamic spectral properties of /uː/ and /ʊ/ in wordlist and conversation styles across 21 Singapore English speakers. The acoustic data were analysed via LMMs and Pillai scores using F1/F2 midpoints, and GAMMs using five points. While static analyses show substantial overlap between the vowels without significant differences across speech styles, dynamic measures reveal nuanced differences in the height and shape of vowel trajectories. The findings contribute to the limited work on Singapore English back vowels and emphasise the need to combine static and dynamic measures to better capture sociophonetic variation in new Englishes.

Index Terms: Singapore English, back vowels, vowel overlap, speech style, individual variation, sociophonetics

1. Introduction

New Englishes or postcolonial Englishes refer to varieties of English spoken in countries with a colonial history and extensive language contact [1]. Singapore English (henceforth SgE) is a notable example, emerging from British colonisation with influences from major local languages such as Hokkien, Malay and Tamil, among others [2]. This sociolinguistic complexity has yielded substantial internal variation within SgE over time. For instance, the term encompasses such named sub-varieties as Singapore Standard English (SSE), Educated Singapore English and Singapore Colloquial English (SCE), distinguished on the basis of formality, functions and intended audience [2, 3], although the delineation between these sub-varieties is not always clear.

Given its diversity, SgE has received considerable attention in the literature [4], particularly regarding synchronic descriptions of its phonological features (e.g. [5-8]). However, sociophonetic research in SgE remains limited [9], with the exception of work on the front vowels /e/ and /æ/ (e.g. [10, 11]), and [12], which examined stylistic patterning of /ɒ/ and /ɔː/. This latter study revealed significant differences across speech styles—wordlist, reading passage, and interview—and age as a significant factor, with speakers aged 21-49 merging /ɒ/ and /ɔː/ and producing both vowels lower in the vowel space, indicative of a sound change in apparent time [12].

The majority of existing studies suggest substantial spectral overlap between /uː/ and /ʊ/ in SgE across different speech styles (e.g. spontaneous vs. read speech) [6, 13, 14] and sub-varieties (SSE vs. SCE [3, 7]). However, some scholars argue for an acoustic distinction between /uː/ and /ʊ/, particularly in SSE spoken in formal contexts (e.g. [3, 7]), although there is a lack of empirical evidence for this distinction. Previous studies also report differences in terms of the position of the two vowels in the vowel space. Both /uː/ and /ʊ/ are described as high back vowels in [6, 7], with /uː/ being spectrally higher but fronter than /ʊ/ [6]. In contrast, [15] suggests /uː/ to be phonetically more back compared to /ʊ/, and [14] observes fronting in both vowels. All this prior work has been based on a static target-based approach to vowel analysis, and none of them specified the exact points from which the formant estimates were taken (i.e. [6, 7, 13, 15]).

The integration of dynamic measures such as generalised additive mixed models (henceforth GAMMs) [16] in sociophonetics has provided valuable new insights into vowel spectral detail and temporal changes [17]. It has been shown that dynamic characteristics of vowels can change diachronically and that listeners attend to dynamicity in the speech signal (see [17]). GAMMs offer a particular advantage in measuring vowel overlap [18], enabling not only statistical comparison of formant trajectories between vowel pairs based on their shape and height, but also of spectral differences between factor groups [19]. Despite the widespread use of GAMMs across varieties of mainstream Englishes, such as American English (e.g. [20]), Australian English (e.g. [17]), British English (e.g. [21]), and New Zealand English (e.g. [22]), their application in New Englishes remains limited, with [9] being the only study on SgE thus far.

The GAMMs employed in [9] analysed trajectory plots and showed little difference between /uː/ and /ʊ/ in F1 over time for a group of ethnic Chinese Singaporean speakers, contrasting with relatively greater variation in F2. Specifically, /uː/ exhibited a backward then forward movement ending close to the starting position, while /ʊ/ showed consistent forward movement in the vowel space, characterised by a lower F2 at the trajectory onset but a higher F2 at the offset. Furthermore, /ʊ/'s F2 differed significantly between the 18-29 and 30-39 age groups, where the 18-29 group produced a spectrally fronter /ʊ/ with a consistently higher F2 trajectory overall. The findings also showed a large amount of interspeaker variation within the younger 18-39 demographic compared to the older 40-69 demographic. These results align with recent sociophonetic research on SgE /ɒ/-/ɔː/ vowels [12] and highlight the role of individual linguistic behaviour within group-level patterns.

Building on the findings of [9], who relied on sentence-read data, this study aims to investigate the impact of style on the acoustic properties of /uː/ and /ʊ/ via both static and dynamic measures. It focuses on young Chinese Singaporeans born after 1987—the year when English was mandated as the only medium of instruction in Singapore [23]. Focusing on controlled versus spontaneous speech, the research questions are: 1) What are the static and dynamic acoustic properties of the /uː/ and /ʊ/ vowels in SgE? 2) How do these properties vary across formal and controlled speech in a wordlist as compared to informal and spontaneous speech in a conversation?

2. Methods

2.1. Participants, procedure and materials

All participants were recorded by the first author in 2023 as part of a larger project investigating the production of monophthongal vowels in SgE. Here we present data from wordlist and conversational speech produced by 21 Chinese Singaporeans (11M, 10F; gender self-identified). By focusing on Chinese Singaporeans, we control for any variation attributable to ethnic differences (see [9]). None of the participants had ever lived in another English-speaking country at the time of data collection and reported no speech disorders. Participants were born between January 1987 and February 2005 (age x̄=26, s=4.57), and their educational backgrounds ranged from junior college and polytechnic to postgraduate levels. 18 participants self-identified English and two (F02, M01) self-identified Mandarin as their first or dominant language, while one participant (F09) listed both English and Mandarin as their dominant languages. Recordings were made using a RODE Wireless GO II dual-channel receiver and transmitter in a quiet room in a public library in Singapore. The audio files were exported in the app, RODE central, and digitised at a sampling rate of 48 kHz with a 24-bit resolution.

The wordlist included five repetitions of each target word (*foot, hood, food, goose*) embedded in the carrier phrase 'He says ____ now'. The spontaneous speech was elicited through conversations between two participants, who had been encouraged to bring a friend and were asked to engage in a casual conversation on a topic(s) provided by the researcher (e.g. *What is your favourite breakfast?*) or a topic(s) of their preference. To ensure comparability between the datasets, only monosyllabic words with a consonant in the syllable coda were extracted from the conversational data (see https://osf.io/ctkym), resulting in a total of 796 analysed tokens (wordlist: 420; conversation: 376). Function words were excluded except for the modal verbs *could*, *should*, and *would.* These generally do not undergo reduction in SgE [8] and were accented in the present dataset.

2.2. Measurements and analysis

All speech files were force-aligned in *WebMAUS* [24], checked and corrected in *Praat* with default settings [25] prior to analysis using emuR [26] in R [27]. The F1/F2 estimates (in Hertz) were automatically extracted across normalised time at 20%, 35%, 50%, 65% and 80% of the vowel duration via *forest* with the parameters set according to speaker gender. Choosing the central portion mitigated the coarticulation effect of the surrounding consonants, while having five data points was visually more informative of the dynamics of the formant trajectories, and is also a prevalent method adopted in recent research (e.g. [28]). All data were visually inspected in EMU, with any datapoints subject to formant tracker errors hand-corrected or removed (~ 2.5%). Lobanov 2.0 normalisation [22] was applied, combining the data for 13 monophthongal vowels across time points, speech styles and speakers, following the assumptions of a vowel-extrinsic formula.

To measure the vowel acoustic properties, static measures such as the Pillai-Bartlett trace, or Pillai score, and Linear Mixed Modelling (LMM) were performed, followed by a dynamic analysis (GAMMs). Pillai scores are one of the conventionally applied approaches to measuring acoustic overlap, with values ranging from 0 (greater overlap) to 1 (greater distinction) [31, 32]. These scores were calculated with *tidyverse* [31] using normalised F1/F2 for group-level behaviour and raw formant values to further examine individual effects (after [34, 35]). LMMs were performed to determine the relationship between the predictors (speech style, vowel, their interaction) and the response variables (normalised F1/F2 estimates) with the inclusion of random effects (speaker and word) using lme4 [34] and lmerTest [35]. Tukey post-hoc tests were performed to locate the source of differences using *emmeans* [36]. GAMMs were used to model and analyse the normalised F1/F2 trajectories using the *mgcv* [37] and *itsadug* [38] packages. These incorporated fixed linear effects between predictors and the response variable (i.e. trajectory height) via parametric analysis, and smooth terms to capture non-linear effects (i.e. trajectory shape). Random smooths extended these smooth functions to random effects by allowing different curves for each value within a grouping variable [19, 21]. The current GAMMs were designed to capture the main effects and interactions; more complex models (e.g. those with item-by-effect random smooth terms) were discarded due to overfitting [39]. Following [21], log-likelihood tests for model comparisons were performed to evaluate predictors involved in interactions. A significant overall comparison was interpreted as an indication that the predictor affected the response variable. Visualisation of model summaries was also used to interpret the model outputs. Autocorrelation was not considered, as the autoregressive error model, which accounts for dependencies between neighbouring data points in the same formant trajectories, did not improve the overall fit (see [28]).

3. Results

3.1. Static results – vowel midpoint

Fig. 1 presents mean normalised F1/F2 plots for all vowels with the wordlist shown in the upper left panel and the conversation in the upper right panel. The lower panel focuses on the target vowels /u:/ and /ʊ/.

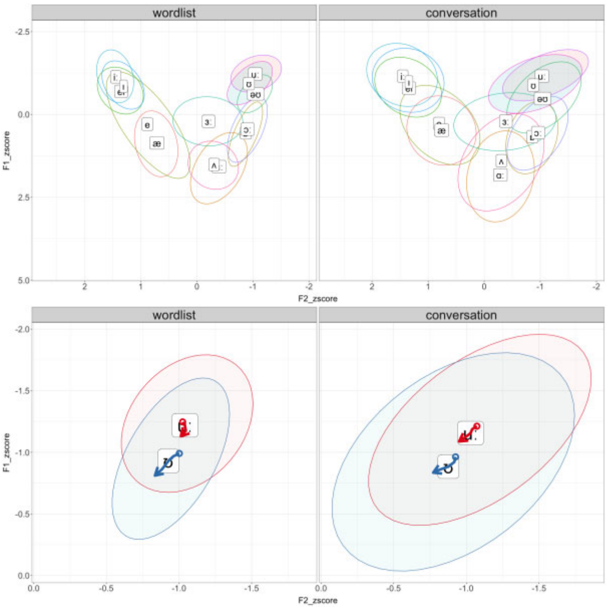


Figure 1: *Normalised F1/F2 estimates for all vowels (top panel) and the vowels /u:/ (red) and /ʊ/ (blue) (bottom panel) by wordlist (left) and conversation (right) with their mean F1/F2 estimates. Ellipses represent 95% confidence intervals; arrows (bottom panel) indicate the direction of the trajectory.*

Neither of the vowels exhibits fronting, as indicated by the position of /u:/ and /ʊ/ in relation to the mid back vowels (Fig.1, upper panel). For both speech styles, /u:/ appears spectrally higher and more back than /ʊ/, as indicated by its mean F1/F2 estimates (Fig.1, lower panel). There is some overlap for the two vowels, with the conversational data showing somewhat greater overlap in the F1/F2 vowel space. Pillai scores corroborated these observations with low scores, indicating substantial overlap between /uː/ and /ʊ/ in both the wordlist (0.11, p<0.001) and the conversation (0.05, p<0.001).

The LMM analysis further supported these findings by reporting non-significant effects of *vowel*, *speech style* and their *interaction* using formant estimates at midpoints, with the exception of *vowel* in F1 (p<0.001). To further explore these results, post-hoc tests were performed, revealing significant differences in F1 between /uː/ and /ʊ/ in both the wordlist (p<0.05) and the conversation (p<0.001). In other words, /uː/ is spectrally higher than /ʊ/ in both speech styles (Fig.1, lower panel). In addition, the random effect of *speaker* emerged as significant for F1 (p<0.05), indicating interspeaker variation in the phonetic realisation of vowels and potentially different patterns across the speakers.

3.2. Dynamic results – vowel trajectories and GAMM

The lower panel of Fig. 1 illustrates the average F1/F2 trajectories for /uː/ and /ʊ/. Both trajectories are short with a similar diagonal gliding (down left) movement across both speech styles. The trajectory for /ʊ/ is slightly longer and shows greater movement compared to /uː/, especially in the wordlist.

Expanding the dynamic analysis further, Fig. 2 shows GAMMs model predictions depicting changes in normalised formant trajectories of /uː/ and /ʊ/ across speech styles, with the baseline model including *vowel*, *speech style*, and their *interaction* as parametric factors, smooth terms over time (five measurement points) and smooth terms over time by *vowel* and by *speech style*, and random smooth terms (*speaker* and *word*).

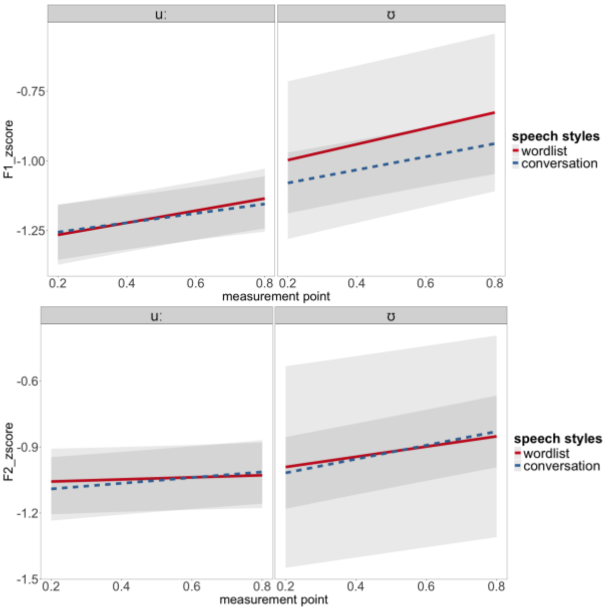


Figure 2: *Model predictions from GAMMs with 95% confidence intervals showing changes in normalised F1 (top) and F2 (bottom) trajectories of /uː/ and /ʊ/ across the wordlist (red) and conversation (blue).*

As illustrated in Fig. 2, /uː/ and /ʊ/ show greater differences in the height of F1 trajectory as compared to F2, with /uː/ exhibiting lower F1 values over time. Both vowels show little difference in F1 slope, but in F2, the almost horizontal flat line of /uː/ suggests little variation over time contrasting with /ʊ/'s positive slope. Greater formant differences are associated with steeper slopes. These observations were confirmed by the GAMMs output (see https://osf.io/paes9). Model comparisons between the full and nested models, which excluded all terms that involved the relevant predictor, revealed *vowel* as the only factor with significant effects for both F1 (vowel: $\chi^2(4) = 4.99$, p<0.05) and F2 (vowel: $\chi^2(4) = 8.75$, p=0.002). Specifically, /uː/ differed significantly from /ʊ/ in the overall height for F1 (determined via a parametric analysis, p<0.01), indicating that /uː/ is spectrally higher than /ʊ/ (i.e. lower F1, see Fig. 2 upper panel), consistent with the results based on the static measures. In addition, /uː/ differed significantly from /ʊ/ in the overall shape for F2 (determined via a non-linear analysis, p<0.001), with /uː/ exhibiting less dynamic movement over time (i.e. a nearly flat line parallel to the x-axis indicating minimal changes, see Fig. 2 lower left panel). In addition, *speaker* and *word* were significant for both formants (random smooth analysis, p<0.001), highlighting speaker-specific behaviour in the production of the two vowels (see §3.3) and lexical-specific effects in the variation of vowel production. Separate GAMMs were fitted to further tease apart acoustic differences between the two vowels in the respective speech styles. /uː/ was produced as a more backed vowel than /ʊ/ in the wordlist (parametric analysis for F2, p<0.01; see Fig. 2 lower panel), but with less dynamic movement over time for both F1 (i.e. non-linear analysis, p<0.05) and F2 (p<0.001). Fewer differences were observed in the conversation, with /uː/ being spectrally higher than /ʊ/ (i.e. F1, parametric analysis, p<0.01; see Fig. 2 upper panel) as the only difference.

3.3. Individual differences

The group-level analysis reported overlap between the two vowels, with /uː/ being significantly higher across the styles (LMMs analysis) and more back than /ʊ/ in the wordlist (GAMMs analysis). However, the significant effects for *speaker* via both LMMs and GAMMs analyses suggested interspeaker variability in vowel productions, which upon further inspection was characterised by three types of behaviour.

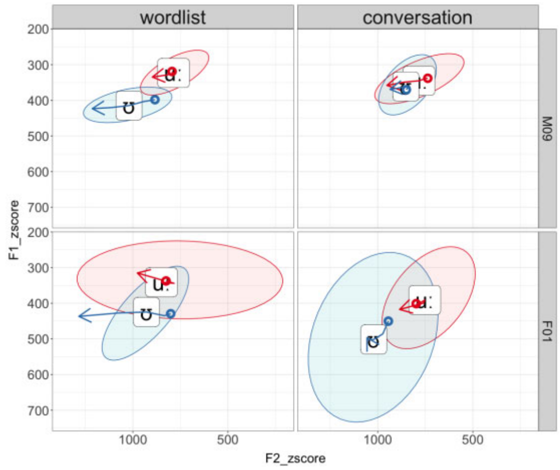


Figure 3. *Normalised F1/F2 estimates of /uː/ (red) and /ʊ/ (blue) for speakers M09 (top panel) and F01 (bottom panel) with their mean F1/F2 estimates. Ellipses represent 95 % confidence intervals; arrows indicate the direction of the trajectory.*

Among the 21 speakers, nine aligned with the group-level patterns (§3.1) and twelve exhibited two distinct behaviours that diverged from the group-level observations. Fig. 3 depicts F1/F2 estimates and trajectory plots for two speakers representative of these distinctive patterns. Speaker M09, whose patterning was also found among five other speakers, exhibited substantial overlap between /uː/ and /ʊ/ in the conversation (Pillai score of 0.13) and almost complete separation in the wordlist (Pillai score of 0.81). In contrast, speaker F01, representative of six other speakers, produced the vowels with a modest overlap in both speech styles and a moderately higher spectral distinction and a greater acoustic distance between the two vowels in the conversation (Pillai score: 0.45 wordlist, 0.59 conversation).

Further, unlike the short trajectories and relatively limited movement across speech styles reported for the vowels in the group-level analysis (Fig. 1, bottom panel), both M09 and F01 produced /ʊ/ with a longer trajectory than /uː/ in the wordlist, and a more dynamic movement overall for /uː/ in the conversation (Fig. 3). The vowels for speaker M09 exhibit uniform leftward trajectories in both speech styles, with a greater trajectory movement for /uː/ in the conversation, while the vowel plot for speaker F01 shows more dynamic changes in /uː/ in the wordlist.

4. Discussion and conclusion

Referring back to the research questions, the analyses based on static and dynamic measures revealed a high degree of overlap between /uː/ and /ʊ/, and this was the pattern for both controlled (formal) and spontaneous (informal) speech styles, with the controlled speech exhibiting moderately less spectral overlap. In particular for RQ1, while the static analysis showed none of the factors as significant except for /uː/ being significantly higher than /ʊ/ (i.e. F1), the dynamic measures revealed *vowel* as a significant factor overall. Specifically, /uː/ was more back and had a less dynamic trajectory over time than /ʊ/ in the wordlist as compared to the conversation.

Similar to [9]'s dynamic observations, our study showed greater variation in F2 for /ʊ/ and less spectral movement for the vowel /uː/, characterised by a flatter trajectory in both speech styles, However, instead of a fronter /uː/ at the onset as in [9], our results suggest that /uː/ starts and ends spectrally higher and more back than /ʊ/ in the vowel space. While these differences may stem from variation in sociolinguistic factors between the two studies (e.g. age), lexical difference, particularly for /uː/, could have contributed to the observed discrepancy. Comparing words from both controlled formal speech styles (i.e. our wordlist data and [9]'s sentence studio-recorded data) revealed that [9]'s three target words—*soon*, *soup* and *food*—primarily featured a /s_#/ environment. Research has shown that preceding coronal phonetic contexts, such as /s_#/, tend to induce greater fronting effect (i.e. higher F2) compared to non-coronal contexts [40]. While an in-depth analysis of word-specific effects is beyond the scope of the current study, differences triggered by words in these phonetic contexts underscore the importance of considering lexical effects in the future study.

RQ2 examines the effects of speech styles, and the group-level results align with previous studies [6, 7, 13, 15] which report substantial overlap between /uː/ and /ʊ/ across speech styles in SgE. However, unlike [14] and [6], the present findings did not observe fronting, with /uː/ being spectrally higher and more back than /ʊ/. Such discrepancies could also come from differences in participant cohorts: our study focused on Chinese Singaporeans rather than Malay Singaporeans [14] or mixed ethnic groups [6], supporting the recent observation about the significant ethnic effect on variation in SgE [9]. Moreover, the present study used vowel midpoints to perform static analysis, and this approach is different from previous studies (i.e. [6, 7, 13, 15]), potentially leading to different acoustic observations.

Individual variation was also present in the current study, similar to the acoustic observations of SgE front vowels reported in [10]. Individual differences were reflected in varied phonetic realisations for both vowels across speech styles: there were instances of overlap in casual speech only, or partial overlap in both speech styles, with the conversational data showing even more distinction than the wordlist. Vowel trajectories also showed differences in shape across individuals.

The distinction between the two vowels in the wordlist but not the conversation across the individuals could be motivated by hyperarticulation, as participants' attention to speech could have increased during a more formal task, such as reading the wordlist [10]. While the inverse acoustic pattern for this in F01 was unexpected, closer examination of the individual Pillai scores and words produced suggests a case of orthographic interference on vowel realisations. The wordlist reading task could have drawn participants' attention to orthography, leading to confusion in pronunciation, particularly for /ʊ/ words resembling /uː/ words with their 'oo' structure (e.g. *foot* versus *food*). The overall forward but limited downward trajectory, indicative of centralisation or fronting, also echoes trends observed in other Englishes (e.g. AmE [20]; BrE [21]).

Given its exploratory nature and a sample restricted to monosyllabic words, the conclusions rendered in this study are subject to further investigation. Moreover, the absence of durational comparisons between the wordlist and conversation may limit our understanding of the acoustic properties of the two vowels. Nonetheless, this study addresses a gap in sociophonetic understanding of SgE back vowels and stylistic and individual effects on phonetic variation and change among young SgE speakers. Furthermore, by using dynamic measures such as GAMMs in a New English, this research makes a significant contribution to the ongoing discussion (e.g. [17, 41]) about using both static and dynamic approaches to examine vowel properties and variability in sociophonetic research on varieties of English. Our future work will extend the examination to polysyllabic words, include duration as an additional factor, and explore the degree of rounding (F3 and the relationship between F2 and F3) through further acoustic analyses.

5. Acknowledgements

We extend our gratitude to Joshua Penny for providing the R script used to generate dynamic vowel plots, and to the two anonymous reviewers for their insightful comments and suggestions.

6. References

[1] Schneider, E. W., Postcolonial English: Varieties around the world. Cambridge: Cambridge University Press, 2007.

[2] Lim, L., Singapore English: A grammatical description, John Benjamins Publishing, 2004.

[3] Low, E.-L., "Chapter 2. Singapore English," in E.-L. Low and A. Hashim [Eds], Varieties of English Around the World, G42: 35–54, John Benjamins Publishing, 2012.

[4] Low, E.-L., "Research on English in Singapore," World Englishes, 33(4):439–457, 2014.

[5] Deterding, D., “Emergent patterns in the vowels of Singapore English,” EWW, 26(2):179–197, 2005.

[6] Deterding, D., “The vowels of the different ethnic groups in Singapore,” in D. Prescott, A. Kirkpatrick, H. Azirah, and I. Martin [Eds], English in Southeast Asia: varieties, literacies and literatures, 2–29, Cambridge Scholars Publishing, 2007.

[7] Lim, L., “Souding Singaporean,” in L. Lim [Ed] Singapore English: A grammatical description, G33:19-56, John Benjamins Publishing, 2004.

[8] Deterding, D., “Phonetics and Phonology,” in Singapore English, 12–39, Edinburgh University Press, 2007.

[9] Low, H. L. C., “Variation and change in the vowels of Singapore English: A sociophonetic study based on the National Speech Corpus,” Nanyang Technological University, Singapore, 2023.

[10] Lan, C. Z., Maxwell, O. and Diskin-Holdaway, C., “Acoustic merger between /e/ and /æ/ in Singapore English: insights into stylistic variation and sub-varietal difference,” Proceedings of the 20th International Congress of Phonetic Sciences, 3661–3665, 2023.

[11] Lan, C. Z., Maxwell, O. and Diskin-Holdaway, C., “An Exploratory Investigation of the /e/-/æ/ and /iː/-/ɪ/ Mergers and Durational Contrasts in Singapore English,” Proceedings of the Eighteenth Australasian International Conference on Speech Science and Technology (SST2022), 191–195, 2022.

[12] Starr, R. L., “Changing Language, Changing Character Types,” in L. Hall-Lew, E. Moore, and R. J. Podesva [Eds], Social Meaning and Linguistic Variation, 315–337, Cambridge University Press, 2021.

[13] Deterding, D., “The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation,” Journal of the International Phonetic Association, 36(2):187–196, 2006.

[14] Tan, R. S. K. and Low, E.-L., “How different are the monophthongs of Malay speakers of Malaysian and Singapore English?,” EWW, 31(2):162–189, 2010.

[15] Deterding, D., “An instrumental study of the monophthong vowels of Singapore English,” EWW, 24(1):1–16, 2003.

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

[17] Cox, F., Penney J. and Palethorpe, S., “Australian English Monophthong Change across 50 Years: Static versus Dynamic Measures,” Languages, 9(3):99, 2024.

[18] Warburton, J., “The Merging of the goat and thought Vowels in Tyneside English: Evidence from Production and Perception,” Newcastle University, 2020.

[19] Sóskuthy, M., “Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction,” 2017.

[20] Stanley, J. A., Renwick, M. E. L., Kuiper, K. I. and Olsen, R. M., “Back Vowel Dynamics and Distinctions in Southern American English,” Journal of English Linguistics, 49(4):389–418, 2021.

[21] Sóskuthy, M., Foulkes, P., Hughes, V. and Haddican, B., “Changing Words and Sounds: The Roles of Different Cognitive Units in Sound Change,” Top Cogn Sci, 10(4): 787–802, 2018.

[22] Brand, J., Hay, J., Clark, L., Watson, K. and Sóskuthy, M., “Systematic co-variation of monophthongs across speakers of New Zealand English,” Journal of Phonetics, 88:101096, 2021.

[23] Pakir, A., “The range and depth of English-knowing bilinguals in Singapore,” World Englishes, 10(2): 167–179, 1991.

[24] Kisler, T., Reichel, U. and Schiel, F., “Multilingual processing of speech via web services,” Computer Speech & Language, 45:326–347, 2017.

[25] Boersma, P. and Weenink, D., “Praat: Doing phonetics by computer.” version 6.3.08, 2023 [Computer program]. Available: http://www.praat.org/

[26] Winkelmann, R., Jänsch, K., Cassidy, S. and Harrington, J., “emuR: Main package of the EMU Speech Database Management System.” R package version 2.4.0, 2023.

[27] R Core Team, “R: A language and environment for statistical computing.” version 4.4.1, 2024 [Computer program]. Available: https://www.r-project.org/

[28] Renwick, M. E. L. and Stanley, J. A., “Modeling dynamic trajectories of front vowels in the American South,” The Journal of the Acoustical Society of America, 147(1): 579–595, 2020.

[29] Nycz, J. and Hall-Lew, L., “Best practices in measuring vowel merger,” The Journal of the Acoustical Society of America, 134(5):4198–4198, 2013.

[30] Heeringa, W. and Van de Velde, H. “A New Vowel Normalization for Sociophonetics,” in Interspeech 2021, 4024–4028, ISCA, 2021.

[31] Wickham, H. et al., “Welcome to the Tidyverse,” JOSS, 4(43):1686, 2019.

[32] Adank, P., Smits, R. and van Hout, R., “A comparison of vowel normalization procedures for language variation research,” The Journal of the Acoustical Society of America, 116(5): 3099–3107, 2004.

[33] Flynn, N., “Comparing Vowel Formant Normalization Procedures,” York Papers in Linguistics Series 2, 11:1–28, 2011.

[34] Bates, D., Mächler, M., Bolker, B. and Walker, S., “Fitting Linear Mixed-Effects Models Using lme4,” J. Stat. Soft., 67(1), 2015.

[35] Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B., “lmerTest Package: Tests in Linear Mixed Effects Models,” J. Stat. Soft., 82(13), 2017.

[36] Lenth, R., “emmeans: Estimated Marginal Means, aka Least-Squares Means.” R package version 1.10.3, 2024.

[37] Wood, S. N., “Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models,” Journal of the Royal Statistical Society Series B: Statistical Methodology, 73(1):3–36, 2011.

[38] van Rij, J., Wieling, M., Baayen, R. H. and van Rijn, H., “Itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.” R package version 2.4.1, 2024.

[39] Sóskuthy, M., “Evaluating generalised additive mixed modelling strategies for dynamic speech analysis,” Journal of Phonetics, 84: 101017, 2021.

[40] Jansen, S. and Mompean, J. A., “GOOSE -fronting in Received Pronunciation across time: A trend study,” Lang Var Change, 35(1):55–77, 2023.

[41] Docherty, G., Gonzalez, S. and Mitchell, N., “Static vs dynamic perspectives on the realization of vowel nuclei in West Australian English,” International Congress of Phonetic Sciences, 2015.

Schwa Optionality in Spontaneous Speech in German: A Meta-Study of Verbal Inflection in Three Corpora

Christine Mooshammer, Marie-Theres Weißgerber, Robert Lange

Institute for German Studies and Linguistics, Humboldt-Universität zu Berlin, Berlin, Germany
christine.mooshammer@hu-berlin.de

Abstract

Standard grammars of German prescribe verb-final schwa in the first person singular for many verbs, e.g. [ha:b@] for *habe* 'have'. This variant, however, is mostly realised without the schwa in spontaneous speech, e.g. [ha:p]. The aim of this study is to investigate the conditions promoting the realisation of schwa in this position. We compare the effect of the segmental context, position within the phrase, speech rhythm, speech rate and situation in three corpora of spontaneous speech. Overall, final schwa was realised only in 21% of all potential cases, with more schwas in IP final position and formal situations.

Index Terms: spontaneous speech, phonetic reduction, German, situational variation

1. Introduction

As has been found in several corpus studies of English, between 25 and 32% of words in spontaneous speech are missing one or more phones compared to the canonical forms (see e.g. [1]). Phonetic reductions are even more common in function words and inflections, and are also found in read speech. As noted by Tucker and Mukai [2], spontaneous speech is a broad term that generally refers to unprepared speech and encompasses a wide range of different speaking styles that vary according to situation, task, addressee etc. The aim of this study is to investigate spontaneous speech phenomena in a wider range of speaking styles (or registers) (see e.g. [3]) and other factors that have been shown to contribute to phonetic variation. In the present study, we focus on the realisation of verb-final schwa inflection in three corpora of spontaneous speech in German. In written German, 1st person singular present tense weak verbs are written with a final <-e> that is pronounced as a schwa [ə], e.g. in *ich mache* 'I make' the canonical pronunciation is [ɪç maχə]. In spontaneous speech, the realisation of the verb-final schwa is optional, and for high frequent verbs, such as *hab-e, seh-e, glaub-e* 'have, see, believe', rather rare (see [4]). In the present tense, dropping the final schwa does not change the meaning (unlike the 3rd person singular in the preterite).

The distribution of this seemingly free variation between e.g. [haːbə] and [haːp] '(I) have' is partly determined by several linguistic and situational factors. First, we will consider how suprasegmental aspects influence reduction phenomena in German. For example, [5] found in several experiments that the speech rhythm, with an alternation between strong and weak syllables, affects whether the final <-e> in adverbs with two alternative spellings, such as *gern, gerne* 'gladly', is realised. Speakers produce schwas more often before stressed syllables in order to avoid stress clashes (see e.g. [6]). Kentner [5], p. 117, identifies a link between the concept of "rhythmic alternation" and the impacts of "stress clash avoidance" and "stress lapse avoidance". A strengthening effect has been found in [7] for following prosodic boundaries with more frequent final schwa in <-en> endings in phrase-final position compared to phrase-medial position. Fast speech rates lead to more frequent reductions, as predicted by Lindbloms H&H Theory [8]. At the segmental level, the following context affects the likelihood of schwas, as found, by Kohler et al. [9] for example, with less frequent final schwas when the following word starts with a vowel. Apart from these phonetic and phonological aspects the semantic load of the inflection also plays a role. As Zimmerer et al. (2014) [10] found for German, reductions of <-t> are less likely when they distinguish between word forms. For verb-final schwa this is the case in the 3rd person singular, preterite, as in *hatte* 'had' which would be homophonic with *hat* 'has' 3rd person singular, present tense without the schwa.

Word frequency, discourse mention, neighbourhood density and predictability also affect how explicit words are pronounced. Clopper & Turnbull (2018) [11] attribute these effects to the smaller processing demands for listeners and speakers of frequent, given and predictable words from high density neighborhoods, which lead to more reduced variants. Based on these factors, they distinguish between 'easy' and 'hard' contexts leading to more or less phonetically reduced forms, respectively. Different speaking styles are also subsumed under this dichotomy by Clopper & Turnbull (2018) [11]. Compared to plain speech, clear speech is easier for listeners to process, but imposes a higher cognitive load on speakers. In the current study, we examine two speaking styles that elicit clear speech phenomena: foreigner-directed and formal speech. Foreigner-directed speech, also known as non-native addressee register (NNAR), is characterised by slower speech rates, hyperarticulation and greater vowel dispersion (see [12] for a review, and e.g. [13]). Definitions of formal speech frequently refer to Labov's notion of "attention given to the speech process" [14]. It is distinctive from informal speech in that speakers often adhere closely to written language [3] and to the standard [15].

In this study, we compare the frequency of verb-final schwa realisations in three corpora of spontaneous German, varying the elicited registers and tasks, as described in Section 2.1. Since NNAR and formal speech are so-called hard registers in the terminology of Clopper and Turnbull (2018) [11], we assume that verb-final schwas are realised more frequently than in informal speech or when speaking to L1 addressees. We also compare free conversations with task-based dialogues (spot the difference task), where the predictions are more exploratory. In terms of processing load for the speaker, the task-based dialogues should elicit more explicit word forms. However, in a previous study [4], we found a slightly but significant increase of verb-final schwas in free conversations compared to task-based dialogues. Furthermore, we will address a subset of the above mentioned factors: phrasal strengthening (schwa adjacent to a boundary), speech rhythm (following word stress), following phonetic context and speech rate. Predictability will be operationalised in a limited context: since word order is rather flexible in German, 1st person singular verb forms are often followed by *ich*, e.g. *habe ich* 'have I', making the co-occurrence of this verb form and the pronoun highly predictable. By comparing these factors in three corpora of spontaneous speech we aim to gain a better understanding of how pronunciation variants are affected by different speaking styles.

2. Method

2.1. Corpora

2.1.1. BeDiaCo: Berlin Dialogue Corpus

The Berlin Dialogue Corpus (BeDiaCo$_{\text{main}}$, [16]) contains spontaneous speech with two different tasks and word lists read aloud. It includes eight face-to-face dialogues by 16 speakers (10 male, 6 female) from Northern Germany between the ages of 18 and 31. Prior to the experiment the participants did not know each other. The two tasks consisted of a free conversation about a topic of their choice for about 15 minutes and to solve two spot-the-difference-tasks, referred to as Diapix tasks (about 8 minutes each). The Diapix task [17] was developed as a dialogue elicitation procedure in which the interlocutors collaborate to find differences between two highly similar pictures without seeing each other's versions. For the recordings, the participants were seated in a sound-attenuated booth and equipped with two headsets from Beyerdynamics (Headset Opus 54). Both microphones were connected to a preamplifier, and one channel was assigned to each speaker.

2.1.2. CoNNAR: Corpus of Non-native Addressee Register

The CoNNAR$_{\text{videocall}}$ corpus was recorded with the aim to elicit foreigner-directed speech with highly proficient learners of German [18]. This sub-corpus has a similar structure to BeDiaCo$_{\text{main}}$. The difference to BeDiaCo is that each participant communicates with an L1 speaker and an L2 speaker, the tasks are the same. 20 German L1 speakers (10 female; age 20–38 years, mean age = 26 years, sd = 4.5 years) were recorded in two sessions with instructed confederates: Once with another German L1 speaker (L1 confederate, n = 4) and once with an English L1 speaker (L2 confederate, n = 4, from the UK and the USA) with self-reported mid to high proficiency in German. Each confederate participated in five sessions with the experimental participants. The order of L1 and L2 confederate experiments was counterbalanced. As in BeDiaCo, the recordings of the 40 experimental sessions consisted of word lists, two Diapix tasks (8 minutes each) and an 8 minute free conversation. Due to the COVID-19 pandemic the interlocutors were placed in neighboring rooms and connected via Zoom. The participants were seated in a sound-attenuated booth, the confederates in an office. The data was recorded using directional microphones (Sennheiser) via Audacity [19] as stereo WAV files. Only the data of the experimental participants are considered here.

2.1.3. RUEG: Research Unit "Emerging Grammars in Language Contact Situations"

The RUEG subcorpus *RUEG-DE German* [20] contains spontaneous speech recordings of native speakers of German. Participants watched a short video of an accident and were asked to summarise the events. The experiment comprises two task set-ups, with one formal and one informal setting, which are intended to yield two different speech registers. The experimental set-ups differed from one another in terms of both the design of the experiment room [20] and the task. In the formal condition, subjects were instructed to report the accident they had just observed to a police officer. In the informal condition, they provided a report of the events to a friend. In both tasks, participants were instructed to imagine the conversation partner and to deliver the report via a voice message. 94 recordings of 47 subjects (25 female, between 13 and 37 years) are analysed.

2.2. Annotation

Table 1: *Potential predictors of schwa realisation*

Predictor	*Levels*
Following context	*Obstruent/ Sonorant/ Vowel*
Stress of the follow. syllable	*Stressed/ Unstressed/ Pause*
Global articulation rate	*Numeric*
Following I	*Yes/ No*
Lemma frequency	*Numeric*

Annotations were carried out automatically and corrected manually in Praat [22] on different tiers (see Figure 1 for RUEG) in all three corpora. One tier contains text transcriptions of the recordings, tokenised on the word-level. An extra tier was inserted to specify the stimulus, which is defined as a first-person singular verb with an optional schwa-suffix. In RUEG data annotations, first-person singular irregular preterite verbs such as "wollte" ('wanted') are included. The third tier contains annotations of the realisation of schwas with binary labels for being present or not. The following context of the potential schwa location was annotated on a different tier: pauses, obstruents, sonorants and vowels. In the case of BeDiaCo and CoNNAR, these annotations were done automatically, and for RUEG manually. The last tier contains labels for the stress of the following syllable, with manually annotated values for unstressed and stressed syllables, based on the auditory impression. In RUEG, the occurrence of "ich" ('I') before or after the target word was annotated manually, and in BeDiaCo as well as CoNNAR automatically extracted by the query function of the emuR database system [23]. The articulation rate was calculated by enumerating the syllables of all instances of the pronunciation-based transliteration per speaker for each task using the R package sylly 0.1-6 [24] with sylly.de 0.1-2 [25] and dividing this value by total length of articulation time. Silent pauses, and extra- and paralinguistic events such as laughing, clicks, and background noises as well as pseudonymised tokens were excluded. Lemma frequency contains the absolute frequency of verbs, centred and logarithmised to the natural base.

2.3. Statistics

We used a binary logistic regression analysis (R packages lme4 [26] and lmerTest [27]) to test which factors influence verb-final schwa realisation. Schwa realisation (reference level with schwa vs. without schwa) was included as the dependent variable, and speakers and lemmas as random intercepts to account for individual and word differences. Each corpus was tested separately. Two sets of models were computed because the predictors following segmental context and stress both include the factor level Pause. Therefore, the first set of models includes task and/or register, stress, centered articulation rate and

Table of Contents for this manuscript

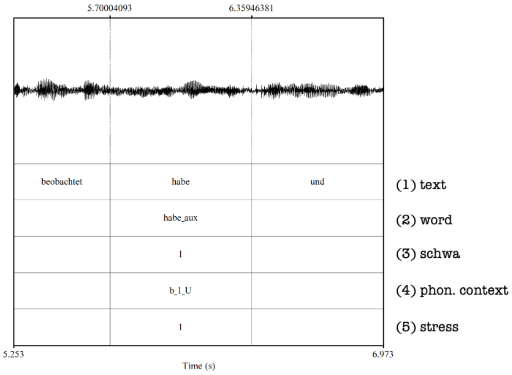


Figure 1: *Annotation layers for the analysis of the RUEG data.*

lemma frequency (see Table 1). For the second set of models, all items with a following pause were deleted and the predictors were following segmental context, following I, task and/or register, centered articulation rate and lemma frequency. By model comparisons it was tested whether inclusion of the interactions improved the models. However, none of the models improved. Significant results are presented in the text. The data and a script with the statistical models can be found at `https://osf.io/hgv9p/`.

3. Results

3.1. Overview

In all three corpora, there are 2624 finite verb forms with a potential final schwa. Only 21% of these are realised. Table 2 gives an overview of the corpora with the number of participants and the amount of available data. The number of data points for RUEG is lower (194) because the recordings per participant are much shorter (74.4 seconds) than for the other corpora (about 48 minutes for CoNNAR and 30 minutes for BeDiaCo). Regarding the percentage of realised schwas (last lines in Table 2), RUEG shows the highest percentage of realised schwas in verb-final position with 33.5%. The lowest percentage of realised schwas (18.3%) is found in the CoNNAR corpus, followed by BeDiaCo with 23.7%.

Table 2: *Overview of corpora.*

	BeDiaCo	CoNNAR	RUEG
Participants N	16	20	47
Duration (min)	180	410	101
Tokens	41036	85949	17538
Tasks	Diapix	Diapix	description
	free conv.	free conv.	–
Channel	face-to-face	video	”imagined”
Addressee	L1	L1	friend
	–	L2	police officer
Potential schwas N	855	1575	194
- with schwa %	23.7	18.3	33.5
- without schwa %	76.3	81.7	66.5

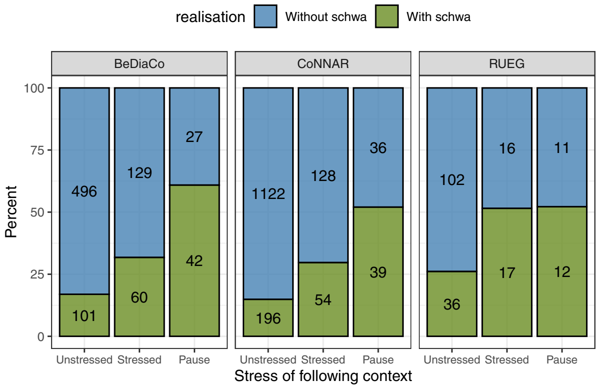


Figure 2: *Effect of prosodic context, percentage and number of tokens.*

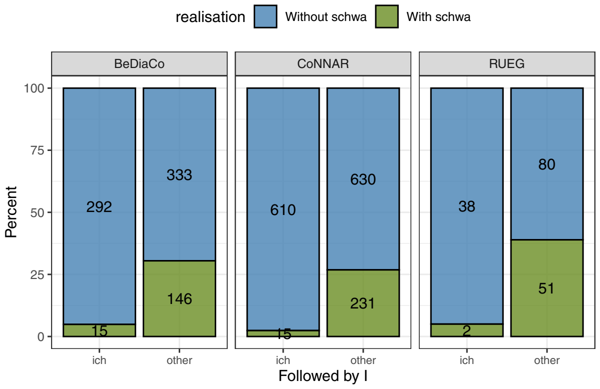


Figure 3: *Followed by ich (excluding following pause).*

3.2. Phrasing, stress and speech rate

Since not all corpora contain prosodic annotations, phrase boundaries are identified by a following pause as a first approximation. Figure 3.2 shows the effect of stress and phrasing. For all three corpora, significantly more schwa are realised before pauses with over 50%, confirming phrase-final strengthening (only compared to unstressed for RUEG). Schwas are more often realised when followed by a stressed syllable compared to following unstressed syllables (significant for BeDiaCo and CoNNAR), which may be a mechanism to avoid stress clashes. Articulation rate has no significant effect on the occurrence of final schwa in all three corpora.

3.3. Frequency and predictability

Lemma frequency is only significant for BeDiaCo with more final schwas in less frequent verbs, as expected. For CoNNAR and RUEG the statistical models do not converge when lemma frequency is included (for the data-set excluding 167 items followed by a pause). Figure 3.3 shows the number of realised schwas depending on whether the verb is followed by ‘I’ or not. Very few schwas (n=32 or 3.3% of the verbs) are realised when followed by ‘I’. This effect is significant for all three corpora.

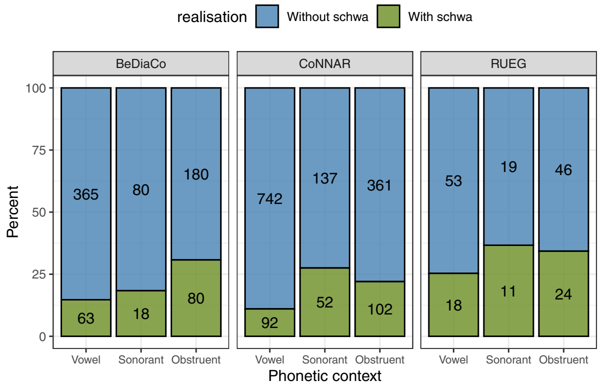


Figure 4: *Effect of the following phonetic context (excluding following pause).*

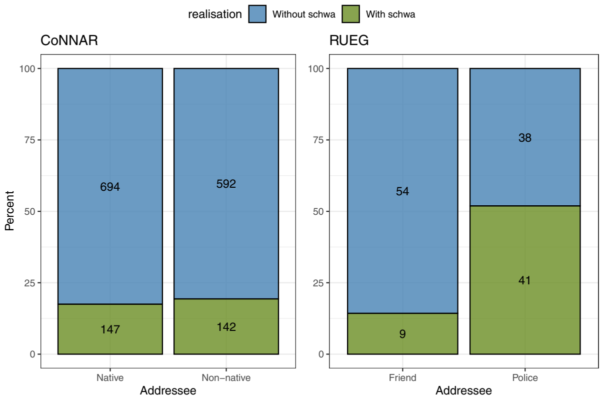


Figure 5: *Addressee comparison in CoNNAR and RUEG.*

3.4. Phonetic context

Significantly fewer schwas are realised when the following word starts with a vowel than with a sonorant for CoNNAR or with an obstruent for BeDiaCo. For RUEG the phonetic context does not show a significant difference, but a similar trend can be observed in Figure 3.4, on the right side.

3.5. Situational variation

Figure 3.5 compares the effect of the addressee on the realisation of schwa for native vs. non-native addressees in CoNNAR (left) and for an imagined friend vs. an imagined police officer in RUEG (right). In CoNNAR the native language of the addressee does not significantly affect the number of schwa realisations whereas in RUEG the more formal situation significantly increases the number of schwa realisations.

4. Discussion

By comparing three corpora, we have identified several consistent factors that affect the realisation of schwa in 1st person singular verb forms in German spontaneous speech. The number of schwa realisations is influenced by phrase boundaries, alternating rhythm, following 'I' and the phonetic context, with more schwa realisations before pauses, stressed syllables, sonorants and obstruents. Another fairly consistent effect is caused by adjacent 'I' with virtually no schwa when followed by 'I'. *Ich* in German is unstressed by default (only 22 *ich* tokens are stressed in our dataset) and starts with a vowel. Excluding all tokens followed by 'I' (37% of all cases) changes the results drastically: for verb forms followed by words other than 'I', the percentage of schwa realisations is not affected by the following stress patterns in all three corpora and even slightly increases for words starting with vowels compared to sonorants and obstru-ents. Therefore, the stress and the phonetic context effects are mainly driven by the very frequent sequence of verbs in the 1st person singular and 'I'. As the pronoun *ich* is the second most frequent word in spontaneous German [28], this co-occurrence is highly predictable and therefore prone to phonetic reduction. As has been suggested in [29] sequences such as *habe ich* 'have I' are contractions or clitics that are resyllabified as in [ha.bɪç]. Note that for verbs with stem-final voiced obstruents, the dele-tion of schwa leads to word-final devoicing in German, e.g. [hap ɪç]. In most cases, however, the resyllabified variant is found in [29], a process similar to 'enchaînement consonantique' in French. We plan to investigate the phonetic details of these con-tractions further.

Another factor influencing phonetic reduction phenomena is speaking style or register. The corpora examined in this study vary in communicative tasks and addressees. As mentioned in the introduction, in [4] we found slightly but significantly more frequent schwa realisations in free conversations than in Diapix tasks. This could not be replicated for the CoNNAR corpus, which generally showed the largest reduction rate (see Table 2). Due to the restrictions during the pandemic, this corpus was not recorded face-to-face in one room, but via Zoom in two adjacent rooms. As Belz et al. 2023 [30] found for read speech, speakers reduce their vowel space in video-conferences compared to co-present situations. They argue that this may be due to the reduced involvement of the speakers in video situations. Contrary to previous findings on NNAR, our participants did not speak more clearly with more frequent final schwa realisations when speaking to non-native speakers. Several reasons could explain this null finding. First, as mentioned above, the speakers were in a video-conference environment. This may have reduced their ability to adapt to their interlocutors. Secondly, the non-native confederates are medium to high-level learners of German with an audible English accent, whereas previous studies have investigated NNAR towards non-native speakers with lower levels of proficiency (see e.g., [12]). The clearest differences between registers are found for formality in the RUEG corpus. Imagining talking to a police officer significantly increased the number of schwa realisations compared to talking to a friend. This striking effect also explains why overall RUEG has the highest percentage of realised schwa. Since these results are based on a subset of the data available in RUEG, we are currently annotating more speakers.

5. Conclusions

Nearly 80% of the possible verb-final schwas are not realised in spontaneous speech in German. This makes variants such as *ich find, ich glaub, ich hab* 'I find, I believe, I have' the standard that should be taught to learners of German. Whether more explicit pronunciations contribute to a foreign accent in German will be investigated in the near future.

6. Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334. We thank Torben Schilling, Miriam Müller, Megumi Terada and Bianca Sell for their annotation work.

7. References

[1] Johnson, K., "Massive reduction in conversational American English". In Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium, 29–54, 2004.

[2] Tucker, B. V. and Mukai, Y., "Spontaneous speech", Cambridge University Press, 2023.

[3] Lüdeling, A., Alexiadou, A. and Adli, A. et al., "Register: Language users' knowledge of situational-functional variation: Frame text of the first phase proposal for the CRC 1412.", in Register Aspects of Language Situation (REALIS), (1), 1-58, 2022.

[4] Lange, R., Sell, B., Terada, M., Belz, M., Mooshammer, C. and Lüdeling, A., "Schwa realisation in verbal inflection in two dialogue registers of German spontaneous speech", in Zeitschrift für Sprachwissenschaft, 1-30, 2024.

[5] Kentner, G., "Schwa optionality and the prosodic shape of words and phrases", in Ulbrich, C., Werth, A. and Wiese, R, Empirical Approaches to the Phonological, 121-151, 2018.

[6] Selkirk, E., "Phonology and Syntax: The Relation between Sound and Structure", Cambridge, MA: MIT Press, 1984.

[7] Niebuhr, O., Görs, K. and Graupe, E., "Speech reduction, intensity, and F0 shape are cues to turn-taking", in Proceedings of the SIGDIAL Conference, 261–269, 2013.

[8] Lindblom, B., "Explaining phonetic variation: A sketch of the H&H theory", in W. J. Hardcastle and A. Marchal [Ed], Dordrecht: Springer, 403-439, 1990.

[9] Kohler, K. J. and Rodgers, J., "Schwa deletion in German read and spontaneous speech", in Spontaneous German speech: Symbolic structures and gestural dynamics, 35: 97-123, 2001.

[10] Zimmerer, F.,Scharinger, M. and Reetz, H, "Phonological and morphological constraints on German/t/-deletions", Journal of Phonetics 45, 64–75, 2014.

[11] Clopper, C. G. and Turnbull, R., "Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors." in Cangemi, F., Clayards, M., Niebuhr, O., Schuppler, B. and Zellers, M., Rethinking Reduction: Interdisciplinary Perspectives on Conditions, Mechanisms, and Domains for Phonetic Variation, 25-72, 2018.

[12] Piazza, G., Martin, c. and Kalashnikova, M., "The Acoustic Features and Didactic Function of Foreigner-Directed Speech: A Scoping Revie" in Journal of Speech, Language, and Hearing Research 65.8, 2896–2918, 2022.

[13] Uther, M., Knoll, M. A., and Burnham, D., "Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech", in Speech communication, 49(1), 2-7, 2007.

[14] Labov, W., "The linguistic variable as a structural unit", New York: ERIC, 1966.

[15] Heylighen, F. and Dewaele, J.-M., "Formality of language: definition, measurement and behavioral determinants", in Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel, 4(1), 1999.

[16] Belz, M. and Mooshammer, C., "Berlin Dialogue Corpus (BeDiaCo) Version 1", Medien-Repositorium Humboldt-Universität zu Berlin, URL: https://rs.cms.hu-berlin.de/phon, accessed on 25 May 2020.

[17] Baker, R., and Hazan, V., "DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs." Behavior research methods, 43, 761-770, 2011.

[18] Lüdeling, A. Mooshammer, Ch., Lange, R., Sell, B. M., & Terada, M., "Corpus of Non-Native Addressee Register (CoNNAR)" Version 1, 2023, URL: https://rs.cms.hu-berlin.de/phon/pages/home.php.

[19] Audacity team, 2020. Audacity (version 2.4.2). https://www.audacityteam.org/

[20] Lüdeling, A. et al. "RUEG Corpus", Version 1.0, https://zenodo.org/records/11234583, 2024.

[21] Wiese, H.. "Language Situations: A method for capturing variation within speakers' repertoires", in Methods in dialectology XVI, 45: 105-117, Frankfurt Main: Peter Lang, 2020.

[22] Boersma, P. and Weenink, D.. "Praat: doing phonetics by computer", Version 6.0.43, http://www.praat.org/, 2023.

[23] Winkelmann, R., Jaensch, K., Cassidy, S. and Harrington, J, "emuR: Main package of the EMU speech database management system", R package version 2.3.0, 2021.

[24] Michalke, M., "sylly.de: Language support for sylly package: German (Version 0.1-2)", 2017, URL: https://github.com/unDocUMeantIt/sylly, accessed on 23 August 2022.

[25] Michalke, M., "sylly: Hyphenation and syllable counting for text analysis (Version 0.1-6)", 2020, URL: https://github.com/unDocUMeantIt/sylly, accessed on 23 August 2022.

[26] Bates, D., Mächler, M., Bolker, B. and Walker S., "Fitting linear mixed-effects models using lme4", in Journal of Statistical Software 67(1), 1–48, 2015.

[27] Kuznetsova, A., Brockhoff, P. and Christensen, R., "lmerTest package: Tests in linear mixed effects models", Journal of Statistical Software 82(13), 1–26, 2017.

[28] Brackhane, F., "Beobachtungen zu Frequenz und Funktionen von ja in deutscher Spontansprache." in Deutsche Sprache 4/2022, 335-363, 2022.

[29] Trouvain, J., Mooshammer, C., Belz, M., Lange, R., "Merging verb forms with "ich" to enchaînement consonantique in German", Book of Abstracts ISSP 2024 - 13th International Seminar on Speech Production, 2024.

[30] Belz, M., Ebert, M., Müller, M., Sun, J., Terada, M., and Xia, Q., "Reduced vowel space in video conferences via Zoom: Evidence from read speech.", JASA Express Letters, 3(10), 2023.

Bilinguals' Segmentation of Unfamiliar Speech Rhythm

Poorani Vijayakumar, Laurence Bruggeman, Mark Antoniou

The MARCS Institute for Brain, Behaviour and Development

poorani.v@westernsydney.edu.au, l.bruggeman@westernsydney.edu.au, m.antoniou@westernsydney.edu.au

Abstract

Listeners rely on their native language rhythm to segment speech. Bilinguals are thought to have certain advantages in linguistic processing, but little is known about their speech segmentation strategies. Here, we investigated bilinguals' segmentation of speech with an unfamiliar rhythm. Monolingual speakers of a stress-timed language (English) and early bilingual speakers of two stress-timed languages (English-Arabic) completed a fragment detection task in a language with an unfamiliar rhythm (Japanese; mora-timed). Preliminary findings showed that the bilinguals did not outperform the monolinguals. Furthermore, neither group applied a moraic segmentation strategy when segmenting Japanese.

Index Terms: language rhythm, speech segmentation, bilingualism

1. Introduction

The regularity of timing units in speech (i.e., the rhythm of a language) tends to influence the way listeners of a language segment speech [1]. In French, for example, words have predictable syllable boundaries, and therefore a regular syllable structure, allowing listeners to use this regularity for speech segmentation [2, 3]. Conversely, English has many possible syllable structures, which makes the syllable an unsuitable unit for segmentation. Rather, English syllables can be classified as strong or weak, and listeners of English may use this regularity to segment speech [4, 5]. In Japanese, the rhythmic structure is based on the mora, which is a unit of syllabic weight. Light syllables consist of one mora, while heavy syllables are bimoraic [1]. The importance of the mora as a rhythmic unit is illustrated in Japanese poetry. A haiku, for instance, has three lines, with five, seven and five morae, respectively, regardless of the number of syllables [1, 6]. Studies on the segmentation of Japanese have found that native listeners of Japanese use a mora-based segmentation strategy for their language [6, 7]. During speech segmentation, listeners leverage whatever rhythm best characterises the phonological regularities of their own native language [1], which makes segmenting the native language very efficient. However, listeners tend to apply their native segmentation strategy to other languages as well, regardless of whether it is an appropriate strategy for that language [8, 9, 10, 11]. This does not have to be a problem, as long as the other language has the same language rhythm as their native language [1, 12]. And sometimes it may even give second-language (L2) listeners an advantage over native listeners, at least in the confines of an experimental task. For instance, Dutch and German L2 listeners of English may use suprasegmental cues of lexical stress in English more accurately than native listeners of that language [9, 10, 11] (but see [13]).

Unfortunately, the speech segmentation literature has thus far mainly focused on monolingual listeners. Consequently, it is as yet unknown how bilinguals, with the knowledge of two languages, might use language rhythm when segmenting speech in familiar and unfamiliar languages. Many bilinguals are flexible language users and switch between their languages seamlessly, while never seeming to compromise on effective communication [14]. Bilinguals tend to develop their languages as required by the environment in which they use them, so that their language knowledge and proficiency are not static and evolve according to necessity and use [15]. During speech comprehension, both of a bilingual's languages are activated and influence one another to varying degrees, so that bilinguals parse speech signals using the linguistic and contextual knowledge of both of their languages [14, 15]. It has been suggested that, perhaps due to this flexibility, bilinguals may possess advantages over monolinguals in several aspects of linguistic processing. Examples of such advantages include metalinguistic awareness [16], phonetic learning [17] and novel word learning [18]. Nevertheless, bilingual advantages do not extend to every aspect of linguistic processing. For example, monolinguals often outperform bilinguals in tasks of speech production, such as naming tasks (e.g., where participants are asked to name drawings), and fluency tasks (e.g., name as many animals starting with the letter F within a limited time) [19].

These mixed findings raise the question whether the 'bilingual advantage' may extend to speech segmentation as well. One study examined a group of balanced early French-English bilinguals, using a French fragment detection task and an English word spotting task [20]. Based on their answer to the question "Suppose you developed a serious disease, and your life could only be saved by a brain operation which would unfortunately have the side effect of removing one of your languages Which language would you choose to keep?" [20, p. 390], the bilinguals were divided into two language dominance groups. Participants who were thus categorised as French-dominant appropriately used a syllable-based segmentation strategy for French, which is syllable timed[1]. However, they also, inappropriately, used this same strategy for stress-timed English. Conversely, the English-dominant listeners used a stress-based segmentation strategy for both languages.

[1] The importance of the rhythmic properties of a language for the strategies its listeners use during segmentation initially led to the assumption that all languages could be categorised into one of three rhythmic classes. French was thus seen as a syllable-timed language, English as stress-timed, while Japanese was classed as a mora-timed language. Although it is now generally agreed upon that this classification oversimplifies a much more complex reality and is therefore considered problematic [e.g., [28]], we opted to use these categorical terms in the present paper in the interest of brevity.

Although the bilinguals were fluent in both of their languages, they appeared unable to switch between their language rhythms and thus resorted to using only one over the other.

In this study, we aim to investigate if being bilingual benefits speech segmentation in an unknown language with an unfamiliar rhythm. If bilinguals have a 'bilingual advantage', we would expect that they might segment speech in an unfamiliar language with an unfamiliar rhythm more efficiently than monolinguals, perhaps even using language-appropriate segmentation strategies. Monolingual speakers of a stress-timed language (English) and early bilingual speakers of two stress-timed languages (English-Arabic) completed a fragment detection task in a language with an unfamiliar rhythm (Japanese; mora-timed). If bilinguals indeed have an advantage over monolinguals, we hypothesised that the bilinguals would detect the target fragments quicker and more accurately than the monolinguals. If the bilinguals use a language-appropriate mora-based segmentation strategy, we also predicted that they would detect the target fragments that were aligned with mora boundaries faster and more accurately than those that were unaligned (see section 2.2 for more details regarding fragment alignment), suggesting a language-appropriate mora-based segmentation strategy. The monolingual listeners were not predicted to respond differently to aligned and unaligned target fragments.

2. Method

2.1 Participants

Participants were 36 Australian English monolinguals (34 female, M_{age} = 26.0, *SD* = 9.3) and 13 English-Arabic bilinguals (10 female, M_{age} = 21.9, *SD* = 7.7), recruited from the participant pool of undergraduate psychology students at Western Sydney University. To assess participants' language experience and proficiency, we used the Language Experience and Proficiency Questionnaire (LEAP-Q) [21]. The bilingual participants acquired English (M_{age} = 3.0; *SD* = 2.9) and Arabic (M_{age} = 3.8; *SD* = 4.0) in early childhood. All but one of the bilinguals reported English as their dominant language. The bilinguals' mean self-reported proficiency was 9.7 out of 10 (*SD* = 0.6) in English and 6.8 out of 10 (*SD* = 1.5) in Arabic. These self-report ratings were confirmed by participants' scores on the Lexical Test for Advanced Learners of English (LexTALE) [22] and the LexArabic [23], with mean scores of 80.8% (*SD* = 6.8) and 65.4% (SD = 12.5) respectively. No participants reported any problems with their speech or hearing, nor any uncorrected vision problems. All participants gave informed consent before completing the experiment. Upon completion of the experiment, participants were reimbursed with course credits.

2.2 Stimuli

Auditory stimuli were taken directly from Experiment 3 of [6] and included eight pairs of meaningful Japanese target words: *tanishi-tanshi; monaka-monka; kanoko-kanko; sanaka-sanka; nanoka-nanka; kinori-kinri; haneda-handa; shinigao-shingao.* Words in each pair had identical initial and final morae and differed only in their medial mora. For example, *tanishi* and *tanshi* both start with the mora *ta* and end with the mora *shi*. The medial mora is *ni* in *tanishi* and *n* in *tanshi*. Pitch accent patterns were matched across the words within in each pair, so that both words were either accented (fall from a high pitch to a low pitch) or unaccented (no fall in pitch). Each target word was combined with two to five additional words to form a word sequence, for a total of 32 word sequences. Each word sequence only consisted of one target word and the rest were filler words. Sequence length varied from three to six words, with the target word always appearing in the penultimate position (e.g., fourth in a sequence with five words). The target fragments that participants were asked to detect always formed part of both target words in a pair, with the mora boundaries of the target fragment aligning with those of only one of the words in the pair. For instance, the fragment *tan* (/ta/ /n/; forward slash indicates a mora boundary) aligns with the morae of the target word *tanshi* (/ta/ /n/ /shi/) and is unaligned with the morae of the target word tanishi (/ta/ /ni/ /shi/). This manipulation was included in the experiment because previous findings have shown that listeners who use mora-based segmentation detect aligned fragments faster and more accurately than fragments that are not aligned [12, 6].

Apart from the 32 word sequences that contained a target word, there were another 32 trials that were catch trials. These did not contain any target fragments and participants were expected to abstain from pressing a button on these trials. The trials were presented in two blocks of 32 word sequences (16 with embedded target words and 16 catch trials with no embedded target words) each. Target words were divided so that one word in each pair appeared in the first block, the other in the second block.

2.3 Procedure

Participants completed all tasks online, using PsychoPy [24] and the Pavlovia platform. The demographics and language background questionnaire was administered via Qualtrics. Participants were instructed to wear headphones throughout the task. Of the monolinguals, 21 reported using in-ear headphones, and 15 wore over-ear headphones while eight of the bilingual participants reported using in-ear headphones, and five wore over-ear headphones. The experiment began with a practice phase to familiarise participants with the task. The participants then completed the fragment detection task. At the start of each trial, participants first saw a fixation point and after 1s were presented with the audio of a target fragment they had to detect (e.g., *tan*). This was then followed by a 3s silence, and then the start of the word sequence (e.g., "*nazo, ekubo, kengaku, **tan**ishi, mamushi*"). Each word in a sequence started 1s after the offset of the previous word. Participants were instructed to press the spacebar as soon as they heard the target fragment within the presented word sequence. Button presses did not terminate trials, and word sequences were played in full for each trial. The following trial then started 3s after the end of the previous trial.

3. Results

Response times (RTs) were measured from the onset of the target fragment. If a button press was recorded after the entire word sequence had finished, it was counted as a miss. Miss rates were calculated by obtaining the percentage of missed responses for each participant, overall and for both the different target alignment conditions. We removed as outliers all trials that had response times faster than 200 ms.

To account for the trade-off between response speed and accuracy (fast button responses can lead to lower accuracy, while high accuracy may come with slower responses) [25], we chose to combine response time and accuracy into a single dependent variable: the inverse efficiency score (IES; [26]). The IES takes into account a participant's RTs as well as their accuracy, and can thus be seen as the RT corrected for the number of errors committed. IES was calculated using the formula IES = (RT/(1-PE)), where RT is the mean response time of correct responses, and PE is the miss rate. For each participant, we calculated one overall IES (based on all experimental trials in the experiment), as well as one IES each for both alignment conditions (i.e., one IES for the trials in which the target fragments aligned with the mora boundaries of the target word, and one IES for the trials with unaligned target fragments). Raw miss rates, raw mean response times, and mean IES are presented in Table 1.

Table 1: *Raw miss rates and raw mean response times for the monolinguals and bilinguals, by condition and overall*

		Condition		**Overall**
		Aligned	Unaligned	
Monolinguals	Miss rates (%)	20.8	12.8	18.8
	RT (ms)	786	1028	783
	IES (ms)	1038	924	997
Bilinguals	Miss rates %	29.5	30.8	29.8
	RT (ms)	981	777	953
	IES (ms)	1123	1595	1567

Figure 1 shows the overall mean IES for the bilinguals and the monolinguals. A comparison between both participant groups using a non-parametric Mann-Whitney U-test in R [27] showed that there were no significant differences between the IES of the monolinguals and bilinguals ($W = 181$, $p = .235$).

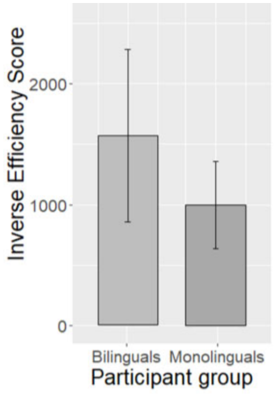


Figure 1: *Bilinguals' (left) and monolinguals' (right) overall IES). Error bars depict standard deviation of the means.*

As mentioned, faster and/or more accurate detection of aligned than unaligned fragments indicates moraic segmentation. Thus, to assess whether participants had used a mora-based segmentation strategy, we further compared the IES for target fragments that aligned with a mora boundary to the IES for targets that did not. Figure 2 shows the IES for monolinguals (left panel) and bilinguals (right panel). A paired-samples Wilcoxon test was used to analyse the IES. Monolinguals showed significantly faster IES ($V = 533$, $p = .001$) for unaligned fragments than aligned fragments. The bilinguals did not show any difference in IES between the types of fragment alignment ($V = 62$, $p = .273$).

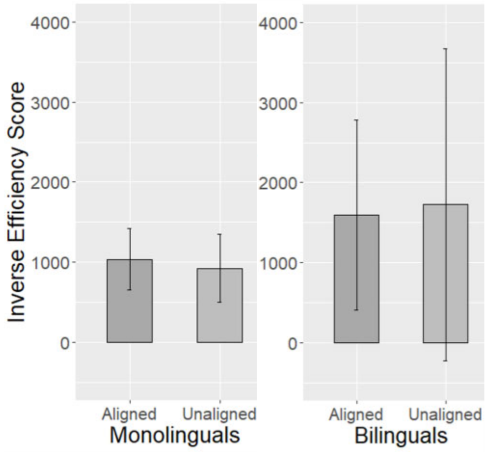


Figure 2: *Monolinguals'(left) and bilinguals' (right) IES by fragment alignment. Error bars depict standard deviation of the mean.*

4. Discussion

The present study investigated how monolingual and bilingual participants segment speech in a language that is unfamiliar to them. Firstly, we had predicted that English-Arabic bilinguals would segment language with an unfamiliar rhythm better than English monolinguals. This hypothesis was not borne out, as bilingual participants were not quicker nor more accurate when detecting the target fragments than the monolinguals. Even though the bilingual listeners possess knowledge of two stress-timed languages (English and Arabic) and the monolinguals know just one stress-timed language (English), this does not seem to provide them with a 'bilingual advantage' when detecting fragments in Japanese, with its unfamiliar mora-timed rhythm.

Our second hypothesis – that bilinguals would use mora-based segmentation and detect targets that align with mora boundaries better than those that do not – was also not supported by our results. This suggests that bilinguals may not have an advantage over monolinguals when it comes to speech segmentation. Our results are in line with previous findings from French-English bilinguals [20] completing similar target fragment detection tasks as the one used here. These bilinguals did not employ an appropriate segmentation strategy in both of their languages. Instead, they relied on only one rhythmic segmentation strategy (i.e., that of their dominant language) for both languages [20]. In contrast to that study, the bilinguals tested here completed the task in a language they did not speak at all. They did not have any knowledge of the appropriate moraic segmentation strategy, and it is therefore perhaps unsurprising that they could not exploit the moraic structure of Japanese during the fragment detection task.

Our results from the English monolinguals in this study showed that they detected unaligned fragments better than aligned fragments. This confirms that the monolingual English listeners did not use a mora-based segmentation strategy, in line with previous studies [e.g., 6]. We had, however, predicted that there would not be any difference in their detection accuracy of these two fragment types, since previous findings with English monolinguals had found no significant differences in the detection speed or accuracy for these types either [6]. On the other hand, other findings have shown that English monolinguals detect fragments faster in words beginning with consonant-vowel-consonant-vowel (CVCV) than in words beginning with consonant-vowel-consonant-consonant (CVCC), regardless of the target fragment itself [3]. This corresponds exactly to the unaligned condition of our study, in which all target words start with CVCV.

This study is a work in progress, and at present may be underpowered due to the small sample of bilingual participants, so it is not possible to draw firm conclusions regarding a potential advantage for bilinguals in speech segmentation. Nevertheless, our preliminary results offer a valuable addition to an understudied research area. Further research is, as always, still needed. The bilinguals in the present study spoke two languages with the same rhythm (stress timing). It would be interesting to see the segmentation strategies used by bilinguals whose languages are from two different rhythmic categories. This could tell us whether knowledge of two different language rhythms may afford bilinguals an advantage when segmenting an unfamiliar language with a rhythm unfamiliar to the bilinguals (e.g., English-Korean bilinguals [stress-timed and syllable-timed, respectively] segmenting mora-timed Japanese). Would this kind of bilingual, unlike the one tested here, employ a language-appropriate segmentation strategy?

In conclusion, this study aimed to investigate if English-Arabic bilinguals detect target word fragments quicker and more accurately in Japanese than English monolinguals. The bilinguals did not segment speech in Japanese, a language unfamiliar to them, better than the monolinguals. Furthermore, neither group applied a mora-based segmentation strategy to Japanese, suggesting that, like the monolinguals, the bilinguals relied on a single segmentation strategy, based on their native language, and do not have an advantage over monolinguals when segmenting speech in an unfamiliar language.

5. References

[1] Cutler, A., "Native listening: Language experience and the recognition of spoken words", The MIT Press, 2012.

[2] Cutler, A., Mehler, J., Norris, D. G. and Seguí, J., “A language-specific comprehension strategy,” Nature., 304:159-160, 1983.

[3] Cutler, A., Mehler, J., Norris, D. G. and Seguí, J., “The syllable’s differing role in the segmentation of French and English.,” Journal of Memory and Language., 25(4):385-400, 1986.

[4] Cutler, A. and Norris, D. G., “The role of strong syllables in segmentation for lexical access,” Journal of Experimental Psychology: Human Perception and Performance.,14(1):113-121, 1988.

[5] Cutler, A. and Butterfield, S., “Rhythmic cues to speech segmentation: Evidence from juncture misperception,” Journal of Memory and Language., 31(2):218-236, 1992.

[6] Otake, T., Hatano, H., Cutler, A. and Mehler, J., “Mora or syllable? Speech segmentation in Japanese,” Journal of Memory and Language., 32:258-278, 1993.

[7] Otake, T., Yoneyama, K., Cutler, A. and van der Lugt, A., “The representation of Japanese moraic nasals,” Journal of the Acoustical Society of America., 100(6): 3831–3842, 1996.

[8] Cutler, A. and Otake, T., “Mora or phoneme? Further evidence for language-specific listening,” Journal of Memory and Language., 33(6):824 – 844, 1994.

[9] Cooper, N., Cutler, A. and Wales, R., “Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners,” Language and Speech., 45:207–228, 2002.

[10] Cutler, A., “Greater sensitivity to prosodic goodness in non-native than in native listeners,” Journal of the Acoustical Society of America., 125(6):3522–3525, 2009.

[11] Yu, J., Mailhammer, R. and Cutler, A., “Vocabulary structure affects word recognition: Evidence from German listeners,” in Proc. Speech Prosody 2020., 2020.

[12] Murty, L., Otake, T. and Cutler, A., “Perceptual tests of rhythmic similarity: I. Mora rhythm,” Language and Speech., 50(1):79-99, 2007.

[13] Bruggeman, L. and Cutler, A., “Listening like a native: Unprofitable procedures need to be discarded,” Bilingualism: Language and Cognition., 26:1093–1102, 2023.

[14] Antoniou, M. “Speech Perception,” in F. Grosjean and K. Byers-Heinlein (Eds.), The Listening Bilingual, Wiley-Blackwell, 43-64, 2018.

[15] Grosjean, F. and Byers-Heinlein, K., “Bilingual Adults and Children: A Short Introduction,” in F. Grosjean and K. Byers-Heinlein (Eds.), The Listening Bilingual, Wiley-Blackwell, 4-24, 2018.

[16] Bialystok, E. “Metalinguistic dimensions of bilingual language proficiency,” in E. Bialystom (Ed.), Language Processing in Bilingual Children, Cambridge, Cambridge University Press, 113-140, 1991.

[17] Antoniou, M., Liang, E., Ettlinger, M. and Wong, P. C. M., “The bilingual advantage in phonetic learning,” Bilingualism: Language and Cognition.,18(4):683–695, 2015.

[18] Kaushanskaya, M. and Marian, V., “The bilingual advantage in novel word learning,” Psychonomic Bulletin & Review.,16:705–710, 2009.

[19] Bialystok, E., Craik, F. and Luk, G. “Cognitive control and lexical access in younger and older bilinguals,” Journal of Experimental Psychology: Learning, Memory, and Cognition., 34(4):859–873, 2008.

[20] Cutler, A., Mehler, J., Norris, D. and Segui, J., “The monolingual nature of speech segmentation by bilinguals,” Cognitive Psychology., 24(3):381–410, 1992.

[21] Kaushanskaya, H., Blumenfeld, K. and Marian, V., “The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later,” Bilingualism: Language and Cognition., 23(5):945-950, 2019.

[22] Lemhöfer, K. and Broersma, M., “Introducing LexTALE: A quick and valid lexical test for advanced learners of English,” Behav Res., 44(2):325-343, 2012.

[23] Alzahrani, A., “LexArabic: A receptive vocabulary size test to estimate Arabic proficiency,” Behavior Research Methods., 56:5529–5556, 2024.

[24] Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E. and Lindeløv, J. "PsychoPy2: Experiments in behavior made easy," Behavior Research Methods., 51:195-203, 2019.

[25] Vandierendonck, A. "A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure," Behav Res, 49:653–673, 2017.

[26] Townsend, J. T. and Ashby, F. G., "Methods of modeling capacity in simple processing systems.," in J. Castellan and F. Restle (Eds.), Cognitive theory, New York, Lawrence Erlbaum Associates,199-239, 1978.

[27] R Core Team, R: A language and environment, Vienna, Austria: R Foundation for Statistical Computing, 2022.

[28] White, L. and Malisz, Z., "Speech rhythm and timing," in C. Gussenhoven and A. Chen (Eds.), Oxford Handbook of Language Prosody, 167-182, Oxford University Press, 2020.

Development of Phonetic Cues in Early L2 Speech Production: The Case of Korean Plosives Pronounced by Native German Speakers

Yeongeun Choi[1], *Christoph Draxler*[2]

[1]Department of Computational Linguistics, University of Zurich; [2]Institute of Phonetics and Speech Processing, Ludwig-Maximilians-University Munich

yeongeun.choi@uzh.ch; draxler@phonetik.uni-muenchen.de

Abstract

This study explores the developmental trends in L2 Korean stop production by German learners, focusing on VOT and F0 of the following vowel. Analyzing results from a reading task with beginners and advanced learners, it was found that beginners demonstrated an L1-like binary stop contrast, relying solely on VOT: short (fortis/lenis) versus long (aspirated). Advanced learners, however, showed notable progress by (1) lengthening VOT for lenis plosives, and (2) employing F0 as a primary cue, with lower F0 after lenis and higher F0 after fortis/aspirated plosives. Thus, learners progressively adapted both 'similar' and 'dissimilar' acoustic cues in a target-like fashion.

Index Terms: Korean stops, Voice Onset Time, fundamental frequency, phonetic cue weighting, L2 speech production

1. Introduction

The most influential second language (L2) speech models highlight the importance of (dis)similarities between learners' first language (L1) and their target language on L2 phone learning. The Speech Learning Model (SLM, [1], [2]) as well as its revised version SLM-r [3] suggest that L2 learners can easily build a new phonetic category for an unfamiliar phoneme that does not exist in their L1 sound system. In the same vein, the Perceptual Assimilation Model (PAM, [4]) predicts that the similarities between L1 and L2 sounds rather hinder the perception of unknown L2 sounds. Accordingly, the PAM-L2 model [5] claims that novice learners tend to substitute L2 phones with the closest L1 equivalent, leading to less precise L2 pronunciation at the initial stage of learning. However, as L2 learning progresses, learners filter informative acoustic features necessary for phonemic distinctions, thereby potentially improving their pronunciation accuracy, which occurs more prominently for dissimilar L2 phones than for similar ones. This study aims to explore these issues further by examining how German learners use suprasegmental cues when producing Korean homorganic plosives, tracking the phonetic shifts that may occur over the course of learning.

Both Korean ([6], [7], [8]) and German ([9], [10]) plosives are voiceless in word-initial position, mainly distinguished along the short-to-long lag *Voice Onset Time* (VOT) spectrum. Despite this similar laryngeal pattern, Korean plosives differ in several respects from the German ones, which may pose notable challenges for German speakers in L2 Korean stop production.

Most importantly, Korean has a phonemic three-way contrast between homorganic plosives as follows: fortis /p* t* k*/[1], lenis /p t k/, and aspirated /p^h t^h k^h/. These plosives are differentiated by two primary acoustic parameters, VOT and the *fundamental frequency* (F0) of the following vowel [7], [11]. Previous studies on sound change of standard (Seoul) Korean plosives (e.g., [8], [12], [13]) have demonstrated that VOT and vowel F0 alternately signal the stop contrasts as primary distinctive cues in word-initial position. The co-existence of these two primary cues has become robust over decades through a VOT merger between lenis and aspirated plosives in standard Korean (for details of VOT merger see also [6]). As a result, VOT serves as a primary cue only for discriminating short-lag VOT (fortis) from long-lag VOT (lenis/aspirated), while vowel F0 serves as a primary cue for signaling a contrast between laryngeal tension (fortis/aspirated) and laryngeal laxness (lenis).

Conversely, standard German has a two-way contrast between lenis /b, d, g/ and fortis /p^h t^h k^h/ plosives, where 'fortis' plosives phonetically resemble Korean aspirated plosives. In German, VOT is the main phonetic cue which distinguishes between homorganic plosives [10], [14], [15], whereas vowel F0 fluctuation [9], [16] occurs unintentionally and naturally as a result of obstruent-intrinsic F0 effects [17].

Drawing from these commonalities and differences between L1 German and L2 Korean in suprasegmental dimensions, as well as their typological mismatch, our research will focus on the following aspects:

(a) Categorization of L2 Korean plosive contrasts at the very early stages of learning. To our knowledge, no studies have been conducted on the production of L2 Korean stops by native German speakers. [18] is the only perception study examining German naïve listeners with Korean stop discrimination. In this test, German listeners exhibited the poorest performance with fortis-lenis pairs, while they effectively discriminated these two plosives from aspirated plosives, respectively. This indicated that German listeners distinguished the three-way contrasts of Korean plosives in a binary manner, in line with their L1 categories. This result is of particular importance, as it suggests that German beginner learners are likely to employ a binary categorization in Korean stop production as well. This expectation may find additional support in other previous research on L2 Korean stops. In both perception and production, learners whose L1 features a two-way stop contrast categorized Korean lenis plosives either with fortis plosives (e.g., English learners [19]) or with aspirated plosives (e.g., Japanese learners [20]), reflecting the stop categorization patterns of their L1. In other words, no learner group employed a two-way discrimination in perception while adopting a three-way discrimination in production. Moreover, these learners predominantly relied on the VOT cue, demonstrating nearly absent proficiency with the F0 cue. This leads us to another expectation that German speakers similarly rely more (or solely) on the VOT cue, which serves as a primary distinctive cue in their L1. We will also analyze whether the incorrectly categorized L2 phonological category – potentially lenis in our case – will be produced as simply equivalent to the single L1 phonological category, or whether it will be

accompanied by good or poorer exemplars (for further details, see PAM [4] and PAM-L2 [5]).

(b) Phonetic shifts in both VOT and F0 cues during the learning process. The developmental changes in the use of VOT and F0 cues will be examined over the course of a one-semester undergraduate course. In terms of the F0 cue, a longitudinal analysis [21] reported that native Mandarin speakers with a tonal language background failed to effectively use the F0 cue, even after approximately one year of Korean language learning, maintaining their conservative behavior in using F0. By contrast, our target group is in the opposite situation, as they do not 'actively' utilize the F0 differences in their L1. Our empirical data will therefore contribute to further insights into how speakers from non-tonal languages manage the F0 cue in L2 stop production.

(c) Change in categorization of L2 plosives at the late early stages. Lastly, the production of advanced beginners will be compared with that of novice beginners, concerning aspects (a) and (b).

To address these research aims, a laboratory speech corpus was collected from 23 German learners of L2 Korean by conducting a carefully designed reading task.

2. Method

2.1. Participants

Two groups of German learners of L2 Korean were recruited in Frankfurt, Germany: 13 beginners and 10 advanced beginners. All participants are female, aged 18-24 years (Mean = 20.3). They attended a Korean language course at Goethe University, three days a week, each session lasting 90 minutes. The 'beginner' group had completed a 10-hour Korean language course, focusing on learning the Korean alphabet and basic pronunciation. The 'advanced beginner' group, hereafter referred to as the 'advanced' group, had completed a 60-hour course, focusing on learning grammar at the sentence level.

2.2. Stimuli

In this study, we adopted a reading task previously conducted in [6] recruiting a different set of participants. The task involved 72 word pairs '**A (target)** and B' displayed within the context of interrogative carrier sentences, which always begin with the adverb *eonje* 'when'. The target word A contains a word-initial plosive, and both A and B are disyllabic words, followed by a postposition *wa/gwa* 'and' for A; an accusative case marker *eul/reul* for B. By employing a word pair framework, we were able to prevent excessive focus on the target word, such as hyperarticulation. The sentences were structured as follows: [*eonje* / A-*(g)wa* / B-*(r)eul* / verb ?]. The test sentences were presented in Korean only.

The word pair items were categorized into four types, by combining them in the following manner: real-real, real-pseudo, pseudo-real, pseudo-pseudo. Pseudo words were used to reduce lexical influence, while simultaneously facilitating the collection of a sufficient sample size. The pseudo A words are quasi-duplicates of the real A words, created by altering the coda of the real A words, for instance, *ttangkong* 'peanut' – *ttanko*; *panmae* 'sale' – *pamae*, as shown in Table 1. By contrast, none of the B words formed minimal pairs with the real and pseudo words. This minimal pair-like speaking setup was designed to analyze learners' pronunciation more precisely, and to enhance the reliability of the speech production data.

Table 1. *Examples of four types of word pair items 'A and B' where the first syllable of the target word A is bold, and the carrier sentence is omitted.*

Type	A (target)	B
real-real	**ttang**kong	hodu
real-pseudo	**pan**mae	donmae
pseudo-real	**ttan**ko	gimbap
pseudo-pseudo	**pa**mae	haneo

To control for vowel-intrinsic effects on F0 [22], the vowels after the target plosives were restricted to the vowels /a/ and /o/. Overall, a total of 1,656 tokens (3 places of articulation x 3 plosive types x 2 vowels x 4 types of word pair x 23 subjects) were collected for analysis. Following the method described by [23], 10 warm-up sentences with other initial consonants were also created, consisting solely of actual Korean words.

2.3. Procedure

The same reading task was conducted in two separate sessions, spaced approximately four months apart. Four speakers participated in both experiments; however, individual variations were not accounted for in this study.

Both experiments followed the same procedure, beginning with the warm-up sentences, followed by the sentences containing the target words. All sentences were randomly presented on a computer screen, using the *SpeechRecorder* software [24]. Generally, speakers read the given sentences aloud a single time. If they mispronounced the target word, they were asked to repeat the entire sentence; however, no hints were provided regarding what was mispronounced.

The participants were recorded in soundproof recording facilities at Goethe University Frankfurt, Germany, using a Røde Lavalier microphone attached to a Zoom H4n audio recorder. All recordings were sampled at 44.1 kHz and 16 bits.

2.4. Measurements and statistical analyses

The collected recordings were automatically segmented and labeled using a Korean forced alignment tool [25]. Subsequently, the time intervals of interest were manually corrected using Praat [26] as follows: for VOT, from the stop release burst to the voicing onset of the following vowel; for vowel F0, from the voicing onset to the offset of the following vowel. These target cues were estimated using separate statistical models as follows.

For the VOT as temporal values, the annotated VOT durations were normalized by calculating z-scores to control for speech rate across speakers. A linear mixed-effects model (LMEM) was conducted in R [27] with z-scored VOT duration as dependent variable, using the R packages *lme4* [28] and *lmerTest* [29]. The model employed as fixed factors *level* (beginner, advanced), *plosive type* (fortis, lenis, aspirated), *place of articulation* (bilabial, alveolar, velar), and *vowel* (/a/, /o/) including their interaction. The random structure included intercepts and random slopes by *speaker* and *target word*. P-values were computed using the Satterthwaite's method with Tukey adjustment for multiple comparisons. Additionally, a post-hoc Tukey's test was performed to examine the level-related VOT differences, using the R package *emmeans* [30].

Prior to statistical analysis on the F0 as non-linear values, each single vowel was divided into 30 time-points in Praat to control for variable vowel durations across speakers and target words. The F0 values measured at each time frame point were normalized using the mean and standard deviation (SD) of the overall F0, averaged across all time frames and all speakers.

The z-scored F0 values were analyzed using generalized additive mixed models (GAMMs) with the R package *mgcv* [31]. This method is particularly effective for examining the non-linear patterns of F0 variation over time-normalized vowel intervals, allowing for the observation of how vowel F0 variations are modulated by the independent variables. The model contained three parametric terms, *level*, *plosive type*, and *vowel*, incorporating by-*target word* random intercepts, along with by-*speaker* random slopes for *Interval*. The overall F0 contours across the learner groups were plotted based on the three-way interaction of *level*, *plosive type*, and *vowel*, including a smooth term for *Interval* by this interaction. In addition, to compare the patterns of vowel F0 among *plosive type* for each group of learners and each vowel type, the parametric coefficients and approximate significance of smooth terms were separately computed considering the interaction between *level* and *vowel*.

3. Results

3.1. Voice onset time

The VOT results revealed significant main effects of *plosive type* (henceforth *type*, *F*[2, 62] = 161.31, *p* < .001) and *place of articulation* (henceforth *PoA*, *F*[2, 62] = 49.18, *p* < .001). In contrast, the effects of *level* (*F*[1, 23] = 0.65, *p* = 0.43) and *vowel* (*F*[1, 62] = 2.24, *p* = 0.14) were not significant. However, the interactions of *level***type* (*F*[2, 51261] = 288.34, *p* < .001), of *level***PoA* (*F*[2, 51260] = 124.35, *p* < .001), and of *level***type***PoA* (*F*[4, 51260] = 46.67, *p* < .001) were confirmed as significant, indicating that learners adjusted VOT durations based on articulatory manner and places of plosives, and altered their use of VOT as they progressed in learning Korean.

Specifically, as depicted in Figure 1, beginner learners distinguished the VOT durations for the three-way plosives in a binary manner: short-lag VOT for fortis/lenis plosives and long-lag VOT for aspirated plosives. Beginner learners showed no significant VOT difference between fortis and lenis plosives, as confirmed by post-hoc pairwise comparisons (*p* = 0.98). In contrast, for advanced learners, the VOT distinction between fortis and lenis plosives became evident (*p* < .001). However, it is noteworthy that the VOT values for lenis plosives are widely distributed, resulting in an overlap of VOT with both fortis and aspirated plosives. Furthermore, VOT durations were generally longer compared to beginner learners.

Considering *PoA*, its significant effect and interaction with *level* and *type* suggest that each learner group displayed different VOT patterns based on *type*. Generally, plosives with a shorter VOT, namely fortis and lenis plosives, were more influenced by *PoA* than aspirated plosives, which have the longest VOT, in terms of VOT adjustment. Commensurate with Figure 1, pairwise comparisons revealed that the VOT durations of fortis and lenis plosives increased from the anterior (bilabial/alveolar) to the more posterior places of articulation (velar). Both bilabial-velar and alveolar-velar comparisons across *level* were significant (*p* < .001). However, within the anterior places of articulation, i.e., between bilabials and alveolars, *PoA* did not significantly affect the VOT for both fortis and lenis plosives. This trend was observed in both the beginner and the advanced group (beginner – fortis: *p* = 0.36; advanced – fortis: *p* = 0.65; lenis: *p* = 0.54). The lenis plosives in beginners exhibited a statistically significant difference between bilabials and alveolars (*p* < .05); however, when compared to their fortis counterparts, the variation in VOT at these two places was not as strong.

The *PoA* effect on VOT of aspirated plosives was not significant in either group: beginner – bilabial-alveolar: *p* = 0.78; bilabial-velar: *p* = 0.10; alveolar-velar: *p* = 0.34; advanced – bilabial-alveolar: *p* = 0.96; bilabial-velar: *p* = 0.26; alveolar-velar: *p* = 0.39).

3.2. F0 contour of the following vowel

The F0 contours are illustrated in Figure 2, incorporating *level*, *type*, and *vowel*, and explaining 72.1% of the variance in the dependent variable. Given the distinctions in F0 contours observed between *level* and between *vowel*, the same model structure was fitted separately to four datasets derived from the interaction of *level***vowel*. Firstly, for the vowel /a/, the models of the beginner and advanced groups explained 68.8% and 76.6% of the variance of *type*, respectively, incorporating random intercepts for *word* and random slopes for *speaker*. The parametric coefficients indicated that in the case of vowel /a/, beginner learners displayed no significant differences in the F0 contour between the three plosives (fortis: β = 0.03, *p* = 0.89; lenis: β = –0.09, *p* = 0.10; aspirated: β = 0.08, *p* = 0.14), nor did advanced learners (fortis: β = –0.34, *p* = 0.27; lenis: β = –0.06, *p* = 0.43; aspirated: β = 0.10, *p* = 0.21). Despite the lack of statistical significance, we can still observe in the top panels of Figure 2 that aspirated plosives before the vowel /a/ consistently exhibited the highest F0 curves at both levels.

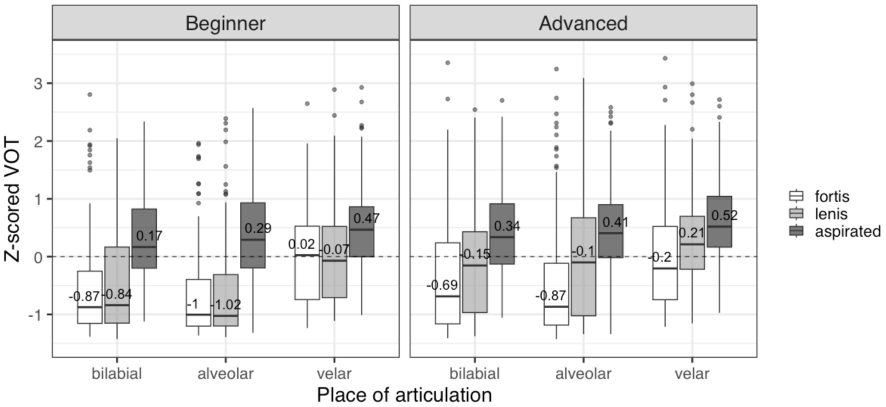


Figure 1: *Z-scored VOT as a function of (a) plosive type and (b) place of articulation across two Korean language levels*

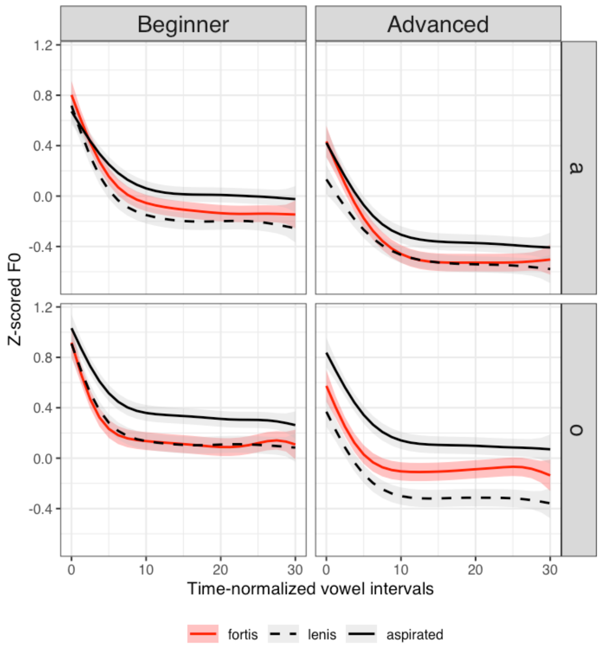


Figure 2: *Average time-normalized standardized F0 contours of the following vowels /a/ and /o/ for fortis, lenis, and aspirated plosives*

For the vowel /o/, the models of beginner and advanced groups explained 66.7% and 75.9% of the variance of *type*, respectively. As can be seen in the bottom panels of Figure 2, both beginner and advanced groups exhibited significantly higher F0 contours for aspirated plosives (beginner: $\beta = 0.22$, $p < .01$; advanced: $\beta = 0.21$, $p < .01$), which was notably distinguished from the other two plosive types. The parametric coefficients demonstrated further that advanced learners significantly discriminate between fortis (red) and lenis (dashed) plosives, using the F0 cue (fortis, intercept: $\beta = –0.04$, $p = 0.90$; lenis: $\beta = –0.20$, $p < .01$) when followed by vowel /o/. However, beginners demonstrated almost identical F0 curves for fortis and lenis (fortis, intercept: $\beta = 0.20$, $p = 0.30$; lenis: $\beta = 0.00$, $p = 0.97$).

Moreover, for advanced learners, the F0 of vowels following lenis plosives shifted downward for both /a/ and /o/, whereas it shifted upward for vowels following fortis plosives for /o/. In comparison with beginner learners, these observations suggest that advanced learners increasingly employed the F0 cue of the following vowel to distinguish between plosive types: laryngeal tension (fortis/aspirated) vs. laryngeal laxness (lenis).

4. Discussion and conclusion

The present study investigated the developmental trajectories of cue use in L2 Korean stop production by German learners during the initial two stages of learning, specifically beginner and advanced beginner levels.

At the initial stage of learning, learners tend to rely solely on the most familiar and reliable acoustic cue, VOT. The results from beginners demonstrate that the three plosives were distinctively categorized into two categories, i.e., those with short-lag VOT (fortis/lenis) and those with long-lag VOT (aspirated). The categorization of lenis plosives as short-lag VOT may be attributed to the orthographic influence being more significant in the initial stage of learning (e.g., [32]) than it is later on. Novice learners, likely influenced more by orthographic information than by the limited auditory inputs they had gathered, appeared to categorize <b d g> as "not (strongly) aspirated" in contrast to <p t k>. These incorrect binary stop categories observed in novice learners also showed a proportional relationship between VOT and vowel F0, suggesting that beginners barely adjust the F0 cue.

The novice learners' L1-like realizations of L2 sounds significantly changed as learning progressed. At later learning stage, specifically after completing a 60-hour Korean course, learners had classified fortis and lenis plosives from a single category into two distinct categories, by adjusting the cue reliance between VOT and vowel F0.

With regard to the VOT adjustment, advanced learners appeared to have acquired during their learning that Korean lenis plosives are more aspirated in word-initial position than they initially recognized, leading to a VOT split between fortis and lenis plosives. However, the VOT durations for both plosives remained widely distributed compared to those of beginners, indicating that the learners' use of VOT has not been fully stabilized yet, and their VOT distribution may continue to change at more advanced stages of learning. Similarly, the overshooting of VOT was also observed across plosive types. Especially, the excessive aspiration found in aspirated plosives may be due to learners' effort to pronounce them with distinctly longer aspiration, particularly longer than that of lenis plosive with lengthened VOT durations. These learners' efforts to precisely control VOT were also highly influenced by articulatory position. The VOT distribution for advanced learners varied according to place of articulation across plosive types: the further back the plosives were articulated, the longer the VOT produced. This contrasts with the articulatory behaviors of (young) native Korean speakers, who exhibit relatively little VOT variation depending on place of articulation [6]. Nonetheless, it is important to note that despite this unstable use of the VOT cue, the direction of changes appears to progress towards a more native-like manner.

The F0 shift of fortis and lenis plosives provides further evidence for the advancement of learners' cue use. Specifically, fortis and lenis plosives did not conform to a universal pattern of obstruent intrinsic effects on F0 (cf. [33]). That is, although the VOT of lenis plosives is as long as that of aspirated plosives, the following vowel exhibits the lowest F0 values, prominently lowering the vowel F0 for lenis plosives. For fortis plosives, by contrast, the VOT was the shortest, yet associated with a higher F0 than vowels following lenis plosives. Considering that F0 variation is an articulatory consequence in L1 German – with longer VOT accompanied by higher F0, our results indicate that German learners have begun adopting the F0 cue as a primary distinctive parameter for discriminating between Korean plosives, showing progress at the advanced beginner level.

Still, this may be a premature conclusion. The significant variation of the F0 cue is strictly interpretated on the basis of our statistical results. The F0 distinction is not as pronounced as that of native Korean speakers (cf. [13]). In addition, the magnitude of the vowel F0 distinction between lenis and fortis plosives fluctuates over vowel intervals depending on vowel types. Thus, future research on more advanced learners in similar contexts may offer further valuable evidence for understanding non-native learners' cue use in L2 speech learning.

5. Acknowledgement

This research was supported by the 2022 Korean Studies Grant Program of the Academy of Korean Studies (AKS-2022-R-061). The authors would like to thank Prof. Dr. Stephan Schmid and Prof. Dr. Eva Maria Luef for their valuable feedback on both the early and revised versions of this paper. We also extend our gratitude to two anonymous reviewers for their insightful comments and suggestions.

6. References

[1] Flege, J. E., “The production of ‘new’ and ‘similar’ phones in a foreign language: Evidence for the effect of equivalence classification”, J. Phon., 15(1), 47–65, 1987.

[2] Flege, J. E., “Second language speech learning: Theory, findings, and problems”, in W. Strange [Ed], Speech perception and linguistic experience: Issues in cross-language research, 92, 233–277, Baltimore: York Press, 1995.

[3] Flege, J. E. and Bohn, O.-S., “The Revised Speech Learning Model (SLM-r)”, in R. Wayland [Ed], Second Language Speech Learning: Theoretical and Empirical Progress, 3–83, Cambridge University Press, 2021.

[4] Best, C. T., “A direct realist view of cross-language speech perception”, Speech Percept. Linguist. Exp., 171–204, 1995.

[5] Best, C. T. and Tyler, M. D., “Nonnative and second-language speech perception: Commonalities and complementarities”, in O.-S. Bohn and M. J. Munro [Eds], Language Learning & Language Teaching, 17, 13–34, Amsterdam: John Benjamins Publishing Company, 2007.

[6] Choi, Y., “Intra- and intersegmental durational compensation of Korean plosives”, in Proc. of the 20th Int. Congress of Phonetic Sciences (ICPhS), 2135–2139, 2023.

[7] Cho, T., Jun, S.-A. and Ladefoged, P., “Acoustic and aerodynamic correlates of Korean stops and fricatives”, J. Phon., 30(2), 193–228, 2002.

[8] Silva, D. J., “Acoustic evidence for the emergence of tonal contrast in contemporary Korean”, Phonology, 23(2), 287–308, 2006.

[9] Jessen, M., Phonetics and phonology of tense and lax obstruents in German, 44, John Benjamins Publishing, 1998.

[10] Kuzla, C. and Ernestus, M., “Prosodic conditioning of phonetic detail in German plosives”, J. Phon., 39(2), 143–155, 2011.

[11] Kim, M.-R. C., “Acoustic characteristics of Korean stops and perception of English stop consonants”, PhD dissertation, The University of Wisconsin-Madison, 1994.

[12] Choi, J., Kim, S. and Cho, T., “An apparent-time study of an ongoing sound change in Seoul Korean: A prosodic account”, PLOS ONE, 15(10), e0240682, 2020.

[13] Kang, Y., “Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: A corpus study”, J. Phon., 45, 76–90, 2014.

[14] Neuhauser, S., “Foreign Accent Imitation and Variation of VOT and Voicing in Plosives”, in Proc. of the 17th Int. Congress of Phonetic Sciences (ICPhS), 1462–1465, 2011.

[15] Kleber, F., “VOT or quantity: What matters more for the voicing contrast in German regional varieties? Results from apparent-time analyses”, J. Phon., 71, 468–486, 2018.

[16] Kirby, J., Kleber, F., Siddins, J. and Harrington, J., “Effects of prosodic prominence on obstruent-intrinsic F0 and VOT in German”, in Proc. of the 10th Int. Conference on Speech Prosody, 210–214, 2020.

[17] Kingston, J., “Segmental influences on F0: Automatic or controlled?”, in C. Gussenhoven and T. Riad [Eds], Tones and Tunes: Experimental Studies in Word and Sentence Prosody, 2, 171–210, Berlin, New York: Mouton de Gruyter, 2007.

[18] Seong, S.-H., “Phonological Transfer and its Hierarchy: L2 Perceptual Acquisition Process of Korean Plosives by German Native Speakers”, J. Korean Lang. Educ., 16(3), 207–226, 2005.

[19] Kim, K.-H., Park, Y. and Chun, Y., “The Production and Perception of the Korean Stops by English Learners”, Speech Sci., 13(4), 51–67, 2006.

[20] Holliday, J. J., “The perception and production of word-initial Korean stops by native speakers of Japanese”, Lang. Speech, 62(3), 494–508, 2019.

[21] Holliday, J. J., “A longitudinal study of the second language acquisition of a three-way stop contrast”, J. Phon., 50, 1–14, 2015.

[22] Whalen, D. H. and Levitt, A. G., “The universality of intrinsic F0 of vowels”, J. Phon., 23(3), 349–366, 1995.

[23] Ladd, D. R. and Schmid, S., “Obstruent voicing effects on F0, but without voicing: Phonetic correlates of Swiss German lenis, fortis, and aspirated stops”, J. Phon., 71, 229–248, 2018.

[24] Draxler, C. and Jänsch, K., “Speechrecorder - a universal platform independent multi-channel audio recording software”, in Proc. of the 4th Int. Conference on Language Resources and Evaluation (LREC), 559–562, 2004.

[25] Yoon, T.-J., “Korean Forced Alignment System”. Online: https://tutorial.tyoon.net/, accessed on 31 Mar 2024.

[26] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer (version 6.3.16)”. Online: http://www.praat.org/, accessed on 31 Aug 2023.

[27] R Core Team, “R: A language and environment for statistical computing (version 4.2.2)”. Online: http://www.R- project.org/, accessed on 26 Dec 2022.

[28] Bates, D., Mächler, M., Bolker, B. and Walker, S., “Fitting Linear Mixed-Effects Models Using lme4”, J. Stat. Softw., 67, 1–48, 2015.

[29] Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B., “lmerTest package: tests in linear mixed effects models”, J. Stat. Softw., 82(13), 2017.

[30] Lenth, R., “emmeans: Estimated marginal means, aka least-squares means. R package (version 1.8.3.)”, 2022.

[31] Wood, S. N., Generalized Additive Models: An Introduction with R, Taylor & Francis, 2017.

[32] Rafat, Y., “Orthography-induced transfer in the production of English-speaking learners of Spanish”, Lang. Learn. J., 44(2), 197–213, 2016.

[33] Ting, C., Clayards, M., Sonderegger, M. and McAuliffe, M., “The cross-linguistic distribution of vowel and consonant intrinsic F0 effects”, PsyArXiv, 2023.

[34] Chang, C. B., “The acoustics of Korean fricatives revisited”, Harv. Stud. Korean Linguist., 12, 137–150, 2008.

[1] The [+tense] feature of Korean fortis consonants is often indicated by the diacritic /*/ (e.g., in [7], [34]).

Articulation of North American English /r/ by Japanese and Mandarin Speakers

Daichi Ishii, Ian Wilson

CLR Phonetics Lab, University of Aizu, Japan

m5281016@u-aizu.ac.jp, wilson@u-aizu.ac.jp

Abstract

Generally, North American English speakers pronounce /r/ as retroflex or bunched, often depending on context and biomechanics. Mandarin speakers also use both articulations when pronouncing Mandarin /r/, but it seems to be speaker-dependent, not context-dependent. This ultrasound study focused on 14 Japanese and 6 Mandarin L2 English speakers pronouncing 138 words containing /r/ in almost all contexts. While Mandarin speakers showed both tongue configurations, all but three Japanese speakers used predominantly retroflex. Those Japanese participants used bunched in almost all contexts and were judged to have near native-like English /r/.

Index Terms: ultrasound, English /r/ articulation, Japanese, Mandarin, retroflex, bunched

1. Introduction

Generally, native North American English speakers pronounce English /r/ in two ways: *retroflex* ([ɻ]) with the tongue tip raised and *bunched* ([ɹ]) with the tongue tip lowered. The characteristic of retroflex [ɻ] is that “the apex is pointed toward the hard palate”, while that of bunched [ɹ] is that “the mid-dorsum of the tongue is raised toward the palate while the lowered apex is retracted from the lower incisors” [1]. Approximately 7% of native English speakers use only retroflex [ɻ], 60% only bunched [ɹ], and 33% use both [2].

Even though there are different ways to produce a North American English /r/, researchers generally agree that listeners do not hear any difference between them [3] [4], even though each seems to use a different pattern of 4^{th} and 5^{th} formants [5].

Although the traditional categorization of North American English /r/ is retroflex versus bunched, other researchers have classified tongue shapes into more categories: front up, tip up, front bunched, mid bunched, and curled up [6] [7]. To a second-language learner, though, such detailed classifications would be confusing. As one goal of our research is to help learners pronounce North American English /r/ better, we focus on whether the tongue tip is raised (retroflex) or lowered (bunched).

Previous studies found that factors affecting the native articulation of /r/ include whether the /r/ is before or after a vowel and the type of adjacent consonant or vowel. Specifically, retroflexion occurs more often in pre-vocalic contexts (Table 1), more often in word-initial or post-labial contexts (Table 2), and more often in back vowel and low vowel contexts (Table 3). Our recent research using all possible contexts (229 words, 5 native North American speakers) supported all those results except for the low vowel preference [8].

One reason for the existence of more than one articulatory strategy for a single acoustic output could be that it is physiologically easier to produce a tip-up or tip-down /r/ depending on the sounds that surround the /r/. Indeed, an articulatory modelling study has shown that reducing the tissue displacement, relative strain, and relative muscle stress does result in widely seen preferences for /r/ articulation by native English speakers [16]. So, if non-native speakers of English could be taught these context-dependent, advantageous /r/ articulations, doing so might help them reduce the effort to articulate English smoothly.

To our knowledge, there are no previous studies focusing on Japanese speakers’ tongue shapes and their distribution during the production of North American English /r/. Japanese has no /r/ or /l/, but has /ɾ/ (a sound with a relatively high tongue tip), and Japanese students are taught to pronounce English /r/ in a retroflex way [17], so we would expect that Japanese speakers of English would use retroflex articulation for English /r/.

Unlike Japanese, the /r/ sound exists in Mandarin Chinese, and Mandarin speakers use three kinds of tongue shapes: retroflex, bunched and post-alveolar [18]. It is not surprising then that Mandarin-English bilingual speakers use both bunched /r/ and retroflex /r/ when speaking English, but they apparently do so in free variation, unlike the pattern of usage employed by native English speakers [19]. In Taiwan Mandarin, people also use both retroflex and bunched /r/ when they speak their native language [20].

In this study, we use ultrasound, a non-invasive method of looking at the distribution of tip-up and tip-down tongue shapes when Japanese and Mandarin speakers of L2 English pronounce North American /r/. We investigate the contexts in which they use one or the other, and whether perceived pronunciation proficiency roughly correlates with that.

2. Method

2.1. Participants

There were 14 native speakers of Japanese (J1–J14; 4 male and 10 female), 3 native speakers of Taiwanese Mandarin (T1–T3; 2 male and 1 female), and 3 native speakers of mainland Chinese Mandarin (C1–C3; 1 male, 2 female). All 20 participants were living in Japan at the time of data collection, 17 computer science students and 3 professors (1 Taiwanese and 2 Japanese). The tongue images for speaker C2 were unclear and so her data were not included.

To evaluate participants’ pronunciation of English /r/, 8 native North-American-English listeners living outside Japan completed an online evaluation task using a Google form. They listened to 84 sound files (4 words × 21 speakers including 2 native North-American-English speakers) including “room” (word-initial /r/), “word” (post-V /r/), “year” (word-final /r/) and “strong” (pre-V /r/) and evaluated participants’ fluency from 1 (“completely non-native”) to 5 (“native-like”). The 84 sound files were arranged in a random order and were presented in

Table 1: *Previous research showing prevocalic /r/ favours retroflexion; including number of participants and words.*

STUDY	TYPE	PARTIC.	WORDS	FAVOURS RETROFLEXION
Delattre & Freeman 1968 [1]	x-ray	46	32	pre-V > post-V
Uldall 1958 [9]	palatography	1	N/A	pre-V > post-V > syl
Hagiwara 1995 [10]	probe-contact	15	6	pre-V > post-V (blade)

Table 2: *Previous research showing contexts favouring retroflexion when /r/ follows a consonant.*

STUDY	TYPE	PARTIC.	WORDS	FAVOURS RETROFLEXION
Delattre & Freeman 1968 [1]	x-ray	46	32	lab > cor > dor
Westbury et al. 1998 [11]	microbeam	53	5	# > lab > dor > /stri/
Guenther et al. 1999 [12]	EMMA	7	5	#, lab > cor > dor
Espy-Wilson & Boyce 1994 [13]	EMMA	1	N/A	other > dor
Tiede et al. 2010 [14]	MRI	4	3-5	cor > other contexts
Uldall 1958 [9]	palatography	1	N/A	cor > other Cs

Table 3: *Previous research showing contexts favouring retroflexion when /r/ is followed by a vowel.*

STUDY	TYPE	PARTIC.	WORDS	FAVOURS RETROFLEXION
Ong & Stone 1998 [15]	ultrasound	1	11	back > front
Tiede et al. 2010 [14]	MRI	4	3-5	low > high

that same order for each listener. The listeners were told that it was North-American English and to specifically rate the "r" sound.

2.2. Stimuli

A list of 138 words containing possible vowel and consonant combinations [(C)rV, Vr(V), and Vr(C)] was created by searching for ARPABET characters in the Carnegie Mellon University (CMU) Pronouncing Dictionary [21]. The number of words for each context is shown in Table 4.

Table 4: *The number of stimuli used for each context. The numbers do not sum to 138 (total words used in experiment) because some words contained more than one context.*

rV		rC	
Context	Words	Context	Words
R + high V	24	R + lab C	15
R + low V	17	R + cor C	33
R + front V	40	R + dor C	11
R + back V	30	R + other C	10
R + high-mid V	9		
R + low-mid V	20		
Vr		**Cr**	
Context	Words	Context	Words
high V + R	14	lab C + R	18
low V + R	15	cor C + R	19
front V + R	18	dor C + R	8
back V + R	37	other C + R	0
high-mid V + R	0		
low-mid V + R	26		

Because Japanese high school students learn about 3,000 English words [22], stimuli were chosen from the Corpus of Contemporary American English (COCA) [23] such that their frequency of occurrence was in the top 3,000. However, some contexts had insufficient words with high enough frequency, so a word commonly known by Japanese was chosen, despite not being in the top 3,000. In one Vr and one Cr context, no words were found meeting the preceding criteria. When choosing stimuli, contexts were balanced in terms of vowels' place of articulation, and types of neighbouring consonants.

In addition to English stimuli, Japanese and Mandarin Chinese stimuli were prepared as their native language stimuli. Japanese stimuli were ら (/ra/), り (/ri/), る (/ru/), れ (/re/) and ろ (/ro/). Mandarin Chinese stimuli were prevocalic rhotic and syllabic rhotic. They were the same stimuli used in [19].

2.3. Apparatus

A Shure Beta 87A microphone and a Steinberg UR22mkII USB Audio Interface were used to record 24 bit, 192 kHz audio. A Famio 8 SSA-530A ultrasound machine with a 3.75 MHz probe was used to record tongue movement. Video was captured and mixed with the audio using a Canopus ADVC-700 Advanced DV Converter and Final Cut Pro on a late 2014 Mac mini computer running macOS 12.7.4. The older Mac was used because it had a built-in FireWire connector compatible with the DV converter. Participants wore a helmet with a 3D-printed probe holder attachment to keep the probe fixed relative to the head.

2.4. Data collection

Firstly, participants filled out their personal background name, age, etc., and signed an agreement allowing us to use their data anonymously. In addition, they filled out a payment form to be paid for their participation. After that, they tried on and adjusted the helmet so that it was snug.

Participants were seated about 2 meters from the laptop screen displaying the stimuli. A microphone test was done to adjust the input volume.

A PowerPoint file containing one of the 138 stimuli per slide, was displayed to participants. The slideshow advanced automatically every two seconds, and after the 46th slide and the 92nd slide, there was a 30-second rest break. The order of the slides was randomized for each participant using a VBA macro. Participants read the 138 English stimuli first and then some stimuli with /r/ or a tap/flap in their native language.

2.5. Data Analysis

Each /r/ frame was extracted from ultrasound movies (.mov) manually. When tongue shapes were somewhat unclear, frames before and after were checked. The frame in which the constriction was the narrowest (highest tongue position) was selected as the /r/ frame and the frame number was noted.

Next, the software "GetContours", which can track tongue contours automatically and also help manual tracking, was used [24] [25]. We set GetContours to provide 100 points along the tongue contours. Firstly, at least three red dots, one of them at the tongue tip, were roughly chosen by clicking along the tongue contour. "Image Forces" helped to make tongue contours clearer for marking those dots by hand. Secondly, "Apply Tracking" was used to make 100 points fit each tongue contour. Sometimes "Apply Tracking" did not work precisely, so manual adjustments were needed. Finally, the points' (x,y) coordinates were extracted into a .tsv file, which was then converted into an .xls file for analysing with Microsoft Excel.

One of the most important analysis steps was deciding how to distinguish categorically between retroflex /r/ (tongue tip up) and bunched /r/ (tongue tip down). We used the following definition: if the slope of the tongue tip for /r/ is higher than the slope of the tongue tip for tap/flap (/r/), the articulation is retroflex and if lower, it is bunched.

The reason for choosing /r/ articulation as the border between retroflex and bunched was that the tongue tip cannot be higher than the alveolar ridge unless it is curled back behind the ridge. Also, /r/ is found in both English and Japanese, so participants' samples were readily available. Each /r/ frame was extracted the same way as extracting the /r/ frames. The sound /r/ was collected from Japanese participants by having them say /ru/, and from Mandarin speakers when they said the second "t" of the English word "strategy" (a tap). The exception were for J10 and J13, so the /r/ in "strategy" was used for them instead.

The tongue tip slopes were calculated based on points 95–100 that had been obtained from GetContours. The 5 slopes between each pair of those 6 points were averaged together for a mean tongue tip slope.

3. Results and Discussion

In the left column of Table 5, the numbers in parentheses indicate the mean native-listener judged proficiency of /r/ production (henceforth, "R-score"), from 1 (completely non-native) to 5 (native-like). Native speakers' R-scores were 4.81 and 4.83. The highest Japanese R-score was 3.94, and 4 participants' scores were higher than 3.00, but the other 10 participants were rated lower than 3.00, meaning their /r/ pronunciation was closer to "completely non-native" than to "native-like". All three Taiwan Mandarin speakers' R-scores were above 3.00, but not as high as the top 3 Japanese speakers. The two mainland Chinese Mandarin speakers' R-scores were both below 3.00. The participants within each native language are listed in R-score order.

Table 5 shows each participant's bunched /r/ rate in each context. Blue indicates contexts in which /r/ was bunched more often than retroflex; pink indicates the opposite. Overall (in "All contexts"), J12, J1, J14, J8, J3, T1, T2, and C1 all use bunched more than retroflex. J1 was the only speaker to use bunched /r/ more in every single context. It was very interesting that J1, who had higher R-score, showed /r/ articulation closest to the tendencies of native speakers (from Tables 1, 2 and 3). Also, J12 and J14 used bunched more in all contexts except for one. This seems to indicate that more native-like pronunciation results from following a native-speaker distribution of bunched/retroflex tongue shapes. An exception, though, was J13, who had the 3rd-highest R-score but used retroflex more in every single context.

The following participants had no blue cells in any context (indicating they used retroflex at least as often as bunched in every context): J13, J11, J10, J5, J6, J9, and T3. Considering the fact that most Japanese participants mainly used retroflex /r/, it is natural to think that they were explicitly taught to pronounce /r/ in a retroflex way [17]. Even without such explicit instructions, it is possible that they chose to use retroflex /r/ naturally on their own because Japanese /r/ is a tap, so the tongue tip is raised.

In contrast to the fact that Japanese mainly used retroflex /r/, Mandarin Chinese speakers were more variable. As shown in [19], low-proficiency Mandarin L2 speakers use retroflex /r/ more in Post-V position than in Pre-V position, and high-proficiency Mandarin L2-English speakers show almost same retroflexion rate in both Pre-V and Post-V position (36% and 30%) . Our results were similar only for high-proficiency Mandarin L2-English speakers. However, one thing that should be noted is that while [19] used standardized English test scores to evaluate fluency, our standard of proficiency was based on R-scores (subjective evaluations from native listeners).

The context in which J1 used bunched /r/ the least in was word-initial and post-labial, but still used it two-thirds of the time. All other Japanese participants except J12 and J14 used retroflex equally or more than bunched /r/ in those contexts. In the post-dorsal consonant context, when native speakers tend to use bunched /r/, J1 used bunched all the time, but all other Japanese speakers except J14 used retroflex most of the time. Japanese participants including J1 (but excluding J12, J13, J14) used bunched /r/ less in Pre-V context than Post-V context, the same as native speakers.

Focusing on the type of vowel following /r/, Table 6 shows that J1 used bunched /r/ 87.5% of the time before high vowels, which have the tongue blade raised but the tip lowered, making a bunched articulation more natural for a neighbouring /r/, as in [14]. Overall, all participants except J13, J14 and C1 used bunched /r/ equally or less often in pre-low than pre-high context. On the other hand, there was not much difference between the rate of bunching for pre-front vowel versus pre-back.

Mandarin speakers who have relatively low R-scores (C1 and C3) used bunched /r/ less in pre-V context than post-V context (like native speakers do). On the other hand, the bunched /r/ rate of T1, T2 and T3 (all with higher R-scores) was almost the same comparing pre-V and post-V. Thus, Mandarin speakers seem to allow for more free variation, unlike many of the Japanese speakers who followed native speaker norms in pre-V versus post-V contexts. Also, although the bunched /r/ rate of T1, T2 and T3 is almost the same, C1 and C3 used bunched /r/ less in Post-C context than Pre-C. The post-C context which

Table 5: *Rate at which /r/ was pronounced with a bunched (tip-down) articulation in various contexts by participants. The "R" number in parentheses is the mean proficiency of /r/ pronunciation from 1 (completely non-native) to 5 (native-like), judged by 8 native North-American-English listeners. Cell background colours are blue if bunched /r/ prevails and pink if retroflex /r/ prevails. Cell shading colours: dodger-blue* ⩾ 90*, deep-sky-blue* ⩾ 70*, light-blue* > 50*, white* = 50*, light-pink* < 50*, hot pink* < 30*, deep pink* < 10.

Partic. (R) Sex	All contexts (%)	Pre-V (%)	Post-V (%)	Pre-C (%)	Post-C (%)	Post-# (%)	Post-Lab. (%)	Post-Cor. (%)	Post-Dor. (%)
J12 (3.94) F	61.4	65.7	60.7	59.5	64.4	66.7	55.6	84.2	37.5
J1 (3.88) M	86.9	78.6	93.4	93.2	77.8	66.7	61.1	84.2	100
J13 (3.69) M	24.2	25.7	21.3	25.7	28.9	16.7	27.8	26.3	37.5
J14 (3.31) F	56.9	65.7	59.0	47.3	57.8	83.3	61.1	52.6	62.5
J11 (2.59) F	24.2	11.4	32.8	31.1	8.9	8.3	0.0	15.8	12.5
J2 (2.53) F	37.9	40.0	47.5	36.5	35.6	16.7	0.0	68.4	37.5
J10 (2.50) F	2.0	1.4	3.3	2.7	2.2	0.0	0.0	5.3	0.0
J7 (2.44) F	36.6	12.9	50.8	58.1	13.3	0.0	5.6	21.1	12.5
J8 (2.34) F	58.8	44.3	73.8	74.3	37.8	41.7	22.2	52.6	37.5
J3 (2.31) F	51.0	34.3	62.3	66.2	33.3	25.0	22.2	52.6	12.5
J5 (2.22) M	7.8	5.7	14.8	9.5	2.2	16.7	0.0	5.3	0.0
J6 (2.22) F	20.9	28.6	27.9	14.9	26.7	8.3	16.7	36.8	25.0
J9 (2.06) F	32.7	15.7	45.9	50.0	15.6	16.7	11.1	15.8	25.0
J4 (1.78) M	46.4	44.3	54.1	50.0	37.8	50.0	50.0	26.3	37.5
T1 (3.63) M	66.0	68.6	65.6	62.2	62.2	83.3	27.8	84.2	87.5
T3 (3.56) F	23.5	22.9	26.2	20.3	17.8	25.0	5.6	31.6	12.5
T2 (3.03) M	51.0	52.9	52.5	47.3	53.3	25.0	38.9	73.7	37.5
C1 (2.81) F	71.9	58.6	83.6	81.1	51.1	91.7	22.2	78.9	50.0
C3 (2.53) M	37.3	17.1	52.5	55.4	13.3	33.3	0.0	21.1	25.0

Table 6: *Rate at which /r/ was pronounced with a bunched articulation in various pre-V contexts by participant. For vowel types, H = high, L = low, F = front, B = back. The R-score numbers in parentheses and the colour coding of cells is the same as in Table 5.*

Partic. (R-score)	H (%)	H-mid (%)	L (%)	L-mid (%)	F (%)	B (%)
J12 (3.94)	70.8	44.4	52.9	80.0	62.5	70.0
J1 (3.88)	87.5	55.6	64.7	90.0	77.5	80.0
J13 (3.69)	16.7	33.3	23.5	35.0	20.0	33.3
J14 (3.31)	58.3	66.7	58.8	80.0	62.5	70.0
J11 (2.59)	16.7	11.1	11.8	5.0	12.5	10.0
J2 (2.53)	50.0	22.2	29.4	45.0	45.0	33.3
J10 (2.50)	4.2	0.0	0.0	0.0	2.5	0.0
J7 (2.44)	25.0	0.0	0.0	15.0	10.0	16.7
J8 (2.34)	54.2	44.4	23.5	50.0	50.0	36.7
J3 (2.31)	29.2	22.2	29.4	50.0	35.0	33.3
J5 (2.22)	4.2	22.2	5.9	0.0	2.5	10.0
J6 (2.22)	45.8	22.2	11.8	25.0	32.5	23.3
J9 (2.06)	29.2	22.2	5.9	5.0	22.5	6.7
J4 (1.78)	58.3	22.2	35.3	45.0	55.0	30.0
T1 (3.63)	83.3	66.6	58.8	60.0	75.0	60.0
T3 (3.56)	45.8	11.1	0.0	20.0	30.0	13.3
T2 (3.03)	70.8	22.2	35.3	60.0	57.5	46.7
C1 (2.81)	41.7	55.6	76.5	65.0	57.5	60.0
C3 (2.53)	20.8	11.2	11.8	20.0	12.5	23.3

favored retroflex /r/ the most among Mandarin speakers was post-labial, similar to some past findings with native speakers (Table 2).

Besides high proficiency Japanese, other Japanese participants also showed similar tendencies to previous research on native speakers, like having a following low or back vowel trigger retroflex /r/. However, their bunched /r/ rate was less than 50% in almost all contexts and the articulatory difference between pre-high versus pre-low or pre-front versus pre-back was not substantial, so it cannot be said that many Japanese speakers use retroflex /r/ and bunched /r/ in the same way as native speakers do.

Based on the results of this research — specifically the fact that participants J12, J1, J14, and T1, who were perceived to be pronounce closer to native North American English /r/, used bunched articulations than retroflex in more contexts — it is tempting to believe that Japanese learners of English should be taught to produce a bunched articulation for /r/. However, research has shown that exclusively teaching bunched articulation to American native English-speaking children (in intervention situations) does not help [26]. What that research *did* find helpful is making sure that learners know *both* varieties so that they can choose the one that best suits their own abilities.

Recall that participants J13 and T3, whose R-scores were higher than many, used retroflex /r/ more often in every context, and that 7% of native speakers use exclusively retroflex /r/ [2], indicating that although bunched articulation may help an L2 speaker's R-score, using bunched /r/ is not a *necessary* condition for pronunciation proficiency.

4. Acknowledgements

We thank the 20 participants for their ultrasound tongue images and the 8 native listeners who judged the /r/ productions. Research was supported by JSPS Kakenhi grant JP20K00606 to the second author.

5. References

[1] P. Delattre and D. C. Freeman, "A dialect study of American r's by x-ray motion picture," *Linguistics*, vol. 6, no. 44, pp. 29–68, 1968.

[2] J. Mielke, A. Baker, and D. Archangeli, "Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/," *Language*, pp. 101–140, 2016.

[3] S. Chen, D. H. Whalen, and P. P. K. Mok, "What R Mandarin Chinese /ɹ/s? – acoustic and articulatory features of Mandarin Chinese rhotics," *Phonetica*, 2024. [Online]. Available: https://doi.org/10.1515/phon-2023-0023

[4] M. Gunji, I. Wilson, and J. Perkins, "Reaction time of Japanese listeners to retroflex and bunched /r/ pronunciation by native English speakers," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3334–3334, 2016.

[5] X. Zhou, C. Y. Espy-Wilson, S. Boyce, M. Tiede, C. Holland, and A. Choe, "A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4466–4481, 2008.

[6] E. Lawson, J. M. Scobbie, and J. Stuart-Smith, "Bunched /r/ promotes vowel merger to schwar: An ultrasound tongue imaging study of Scottish sociophonetic variation," *Journal of Phonetics*, vol. 41, no. 3-4, pp. 198–210, 2013.

[7] H. King and E. Ferragne, "Loose lips and tongue tips: The central role of the /r/-typical labial gesture in Anglo-English," *Journal of Phonetics*, vol. 80, p. 100978, 2020.

[8] D. Ishii and I. Wilson, "Contexts and biomechanics of native English speakers' /r/ tongue shapes," in *Proceedings of the 2024 Spring Meeting of the Acoustical Society of Japan*. Acoustic Society of Japan, 2024, pp. 1023–1026.

[9] E. Uldall, "American 'molar' R and 'flapped' T," *Revista do Laboratôrio de Fonética Experimental da Faculdade de Letras da Universidade de Coimbra*, vol. 4, pp. 103–106, 1958.

[10] R. E. Hagiwara, *Acoustic realizations of American /r/ as produced by women and men*. UCLA Working Papers in Phonetics, vol. 90, 1995.

[11] J. R. Westbury, M. Hashi, and M. J. Lindstrom, "Differences among speakers in lingual articulation for American English /ɹ/," *Speech Communication*, vol. 26, no. 3, pp. 203–226, 1998.

[12] F. H. Guenther, C. Y. Espy-Wilson, S. E. Boyce, M. L. Matthies, M. Zandipour, and J. S. Perkell, "Articulatory tradeoffs reduce acoustic variability during American English /r/ production," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2854–2865, 1999.

[13] C. Espy-Wilson and S. Boyce, "Acoustic differences between "bunched" and "retroflex" variants of American English /r/," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2823–2823, 1994.

[14] M. K. Tiede, S. E. Boyce, C. Y. Espy-Wilson, and V. L. Gracco, "Variability of North American English /r/ production in response to palatal perturbation," in *Speech Motor Control: New developments in basic and applied research*, B. Maassen and P. van Lieshout, Eds. Oxford University Press, 2010, pp. 53–67.

[15] D. Ong and M. Stone, "Three-dimensional vocal tract shapes in /r/ and /l/: A study of MRI, ultrasound, electropalatography, and acoustics," *Phonoscope*, vol. 1, no. 1, pp. 1–13, 1998.

[16] I. Stavness, B. Gick, D. Derrick, and S. Fels, "Biomechanical modeling of English /r/ variants," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. EL355–EL360, 2012.

[17] N. Shusse, "English Sounds: Learning and Teaching," *Kokugakuin Zasshi*, vol. 118, no. 6, pp. 1–12, 2017. [Online]. Available: https://k-rain.repo.nii.ac.jp/records/288

[18] K. Xing, "Phonetic and phonological perspectives on rhoticity in Mandarin," Ph.D. dissertation, The University of Manchester, 2021.

[19] S. Chen, D. H. Whalen, and P. P. K. Mok, "Production of the English /ɹ/ by Mandarin–English bilingual speakers," *Language and Speech*, vol. 0, no. 0, p. 00238309241230895, 2024. [Online]. Available: https://doi.org/10.1177/00238309241230895

[20] Y.-h. S. Chang, "Articulation of the rhotic /ɹ/ in Taiwan Mandarin," in *Ultrafest XI: Extended Abstracts*, I. Wilson, A. Mizoguchi, J. Perkins, J. Villegas, and N. Yamane, Eds. University of Aizu, 2024, pp. 118–119. [Online]. Available: https://doi.org/10.5281/zenodo.12578650

[21] K. Lanzo, "The CMU Pronouncing Dictionary," accessed Jan., 18, 2024. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[22] Ministry of Education, Culture, Sports, Science and Technology (MEXT), "Explanation of Courses of Study for Senior High Schools Foreign Language Edition English Edition," 2010, https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2010/01/29/1282000_9.pdf.

[23] M. Davies, "One-billion word Corpus of Contemporary American English (COCA), 1990–2019," 2020, https://www.english-corpora.org/coca/.

[24] M. Tiede and D. Whalen, "GetContours: an interactive tongue surface extraction tool," *Proceedings of Ultrafest VII*, 2015.

[25] M. Tiede, "GetContours: Tongue contour fitting software; v3.5. [Computer program]," 2021, https://github.com/mktiede/GetContours.

[26] T. M. Byun, E. R. Hitchcock, and M. T. Swartz, "Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 6, pp. 2116–2130, 2014.

Characterizing Rhotic Articulation in Australian English using Ultrasound

Michael Proctor, Jae-Hyun Kim, Joshua Penney,
Louise Ratko, Felicity Cox

Department of Linguistics, Macquarie University, Australia
michael.proctor@mq.edu.au

Abstract

Rhotic approximants in English are produced with considerable articulatory variability across speakers and contexts. Although this variability has been examined in North American, British and New Zealand varieties, it has not been documented using instrumental methods in Australian English (AusE). We therefore examined rhotic approximants produced by six speakers of AusE in nine vowel contexts using ultrasound tongue imaging. Three broad patterns of tongue shaping were observed, and speakers differed in the type and degree of vocalic influence on rhotic posture. Implications for phonological characterization of AusE /ɹ/ are discussed, along with methodological considerations for ultrasound studies.

Index Terms: rhotics, ultrasound tongue imaging, Australian English, liquids, articulation

1. Introduction

The phonetic and phonological characterization of rhotic approximants remains an ongoing topic of research, and the complexity and variability of this class of sounds presents many challenges.

Acoustically, many rhotic approximants are characterized by a lowered third formant in some environments [1, 2], but F3 trajectories are not a robust universal perceptual correlate of rhoticity, and acoustic properties of /ɹ/ exhibit complex variability across speakers and environments [3]. Furthermore, the relationships between acoustic properties and rhotic articulation are not well understood [4, 5]

American English /ɹ/ is produced with variable tongue configurations, which have been described using taxonomies that vary in complexity describing two [6], six [7], and eight [8] different categories of rhotic articulation. The most common defining characteristics proposed to differentiate these variant /ɹ/ types in American English – and similar rhotic variation in other Englishes – include ‘retroflex’ vs. ‘bunched’ [9], or ‘tip-up’ vs. ‘tip-down’ [10]. The distinction is complicated by individual speaker variation, sound change, and inconsistent percepts of rhotic type. In an ultrasound study of 27 speakers of American English, Mielke et al. found that “two speakers used only retroflex /ɹ/, sixteen use only bunched /ɹ/, and nine use both /ɹ/ types, with idiosyncratic allophonic distributions”. Furthermore, “these allophony patterns are covert, because the difference between bunched and retroflex /ɹ/ is not readily perceived by listeners” [11, p.101].

Further insights have been provided into English /ɹ/ through analysis of the gestures involved in production. Articulatory studies of North American Englishes using MRI [12, 13], X-ray microbeam [14, 15], ultrasound and video [16, 15, 17] have revealed that /ɹ/ involves coordination of two lingual gestures, with an additional labial gesture observed in onset environments [18]. Australian English (AusE) is a non-rhotic variety [19], so although the properties of rhotics occurring in coda environments are not relevant, the phonetic similarities of AusE onset /ɹ/ suggests that it involves similar goals of production as those of other English varieties.

Heyne et al. [20] investigated rhotic articulation in New Zealand English (NZE), which like AusE, is a non-rhotic variety. Sixty-two speakers produced 13 words containing /ɹ/ in different phonological environments, and tongue shapes were categorized according to the four main patterns described by Delattre and Freeman [7]. 25 speakers consistently produced tip-down variants, 12 consistently produced tip-up rhotics, and 25 speakers produced /ɹ/ with variable tongue shapes; tip-up allophones were more commonly observed in back vowel contexts, and tip-down before high front vowels.

To date, no systematic study of AusE /ɹ/ articulation has been conducted; however, a pilot ultrasound study of rhotics in six speakers from Sydney [21] reveals patterns of production similar to the NZE study [20]. Four broad tongue shapes were observed in word-initial /ɹ/ produced before three different vowel qualities /iː-ɐː-oː/, but further details of production were not analyzed.

Liquid consonants, /ɹ/ and /l/, are typically amongst the most difficult English sounds to master for both first [22] and second language learners, and studies of phonological acquisition in monolingual AusE speaking children in Sydney show that lateral approximants are acquired relatively late [23]. These developmental trajectories suggest that Australian English /ɹ/ may also be characterized by gestural complexity requiring fine control of articulators [24], the details of which are not yet understood. More data is needed to understand the goals of production of AusE /ɹ/ and the articulatory variability it exhibits across speakers and environments.

1.1. Aims

The aim of this study is to investigate lingual articulation of /ɹ/ in AusE, using ultrasound tongue imaging. To the best of our knowledge, this is the first instrumental study to examine AusE rhotic production in a wide range of vowel contexts, allowing us to offer initial insights into patterns of /ɹ/ production. A secondary aim of this study is to establish robust methods and experimental materials for ongoing studies of English rhotics in Australia.

2. Methods

Six female adult monolingual speakers of AusE participated in the study (Mean age = 23.2 years, SD = 2.9, range: 19–27). All participants were born and raised in Australia and had at least one parent who was also born and raised in Australia. All participants had completed their primary and secondary education in Australia. Data were acquired as part of an ongoing study of AusE rhotics. Participants were undergraduate students at Macquarie University, and received course credit for their participation. The ethical aspects of this study have been approved by the Human Research Ethics Committee of the institution affiliated with the authors.

2.1. Experimental Materials

Rhotics were elicited in word-initial position before nine different vowels distributed across three broad places of articulation (Table 1). Each target word was elicited in the carrier phrase *‘It’s a ___’* to ensure that the tongue body was in a neutral position for schwa prior to the rhotic production, enabling observation of both retraction and raising for the following onset rhotic. Each individual item was presented orthographically on a monitor in a pseudo-random order and read aloud by the participant in a self-paced recording session divided into three blocks. During the recording session, participants also produced rhotics in a range of other contexts, which are not analysed here. The elicitation was monitored by the experimenter so that trials compromised by mispronunciations, atypical prosody or noise interference could be re-recorded immediately. A total of 9 (items) × 3 (repetitions) × 6 (participants) = 162 trials were recorded for analysis.

Table 1: *Stimuli used to elicit Australian English /ɹ/ in word-initial position before nine different vowels.*

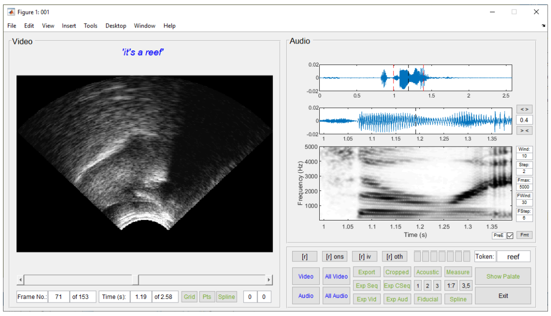
Item	Target	V Place
‘It’s a reef’	/ɹiːf/	High Front
‘It’s a rip’	/ɹɪp/	High Front
‘It’s a ref’	/ɹep/	High Front
‘It’s a rap’	/ɹæp/	Low
‘It’s a raft’	/ɹɐːft/	Low
‘It’s a rough’	/ɹɐf/	Low
‘It’s a raw’	/ɹoː/	Back
‘It’s a rook’	/ɹʊk/	Back
‘It’s a rob’	/ɹɔb/	Back

2.2. Data acquisition

Lingual articulation was tracked in the midsagittal plane using ultrasound tongue imaging. Data were elicited and recorded using the Articulate Assistant Advanced (AAA) software Version 220.2.0 [25] at Macquarie University. A microconvex probe (2-4 MHz, 20 mm radius) was located beneath the participant’s chin, held in place using an Articulate Instruments Aluminium Probe Stabilisation Headset [26].

Ultrasound video data were acquired at a temporal resolution of 55 to 60 f.p.s. with the probe frequency at 3 MHz, depth at 120mm, the focus depth at 96 mm and with a 83.2° field of view. Speech audio was recorded concurrently at a sampling rate of 22,050 Hz using a RØDE NTG1 condenser shotgun microphone located approximately 30 cm in front of the participant, offset 15°, connected to a Focusrite Scarlett Solo 3rd Generation preamplifier.

Hard palates were located using water swallow trials and by fitting a convex hull to the region of maximum lingual excursion observed during obstruent production [27]. Palates were manually traced using the fan spline in the AAA system. Midsagittal palate traces were exported as a 12-point set of cartesian coordinate pairs defined with respect to a fiducial line located immediately above the ultrasound probe.



Frame No.: 71 of 153 Time (s): 1.19 of 2.58 0 0

Token: reef

Figure 1: ***Ultrasound Inspection and Analysis***. *Matlab-based tool used for inspecting and analysing time-aligned ultrasound video and audio data.*

Video data were exported from the AAA system in uncompressed AVI format (RGB24 encoding) with a spatial resolution of 400 × 300 px over a 188 × 141 mm field of view (1 px = 0.47 mm). Companion audio recordings were exported in uncompressed 16 bit mono WAV format.

2.3. Phonetic data analysis

Ultrasound video and companion audio recordings were inspected using a custom Matlab-based graphical user interface facilitating frame-by frame navigation of exported AAA data with time-aligned audio (Fig. 1).

Rhotic targets were identified by inspecting ultrasound video sequences in the interval following the schwa. Two dimensions of rhotic articulation were tracked: (i) raising/advancement of the coronal part of the tongue towards the palate or alveolar ridge, and (ii) retraction of the posterior part of the tongue dorsum observable in the region anterior to the hyoid shadow. The rhotic target was identified as the frame in which the coronal part of the tongue achieved maximal excursion away from the initial lingual posture. Where the coronal part of the tongue maintained a maximally raised/advanced posture for multiple frames, the central frame was chosen.

Three phonetically trained analysts independently identified the rhotic target frame in each utterance in the experimental corpus. 486 target frame numbers were recorded (3 analysts × 162 trials) and compared. Analysts agreed on rhotic target frames for most trials, and in no case did target frame numbers differ by more than 4 across analysts; in these cases, ultrasound video data were re-examined to check rhotic targets.

2.4. Ultrasound image analysis

Image frames at each rhotic target were exported in uncompressed JPEG format. The three frames corresponding to rhotic targets for each repetition in each vowel context were combined into a single image by computing the mean intensity of each pixel in the 400 × 300 matrix. 27 images were generated for each participant, illustrating mean tongue posture for each rhotic in each vowel context. Images were further combined to illustrate rhotic tongue posture at each broad place of articulation of context vowels by generating mean images from the set of all target frames for [iː-ɪ-e] (front), [æ-ɐː-ɐ] (low), and [oː-ʊ-ɔ] (back) vowel contexts. Palate traces were superimposed on images for reference.

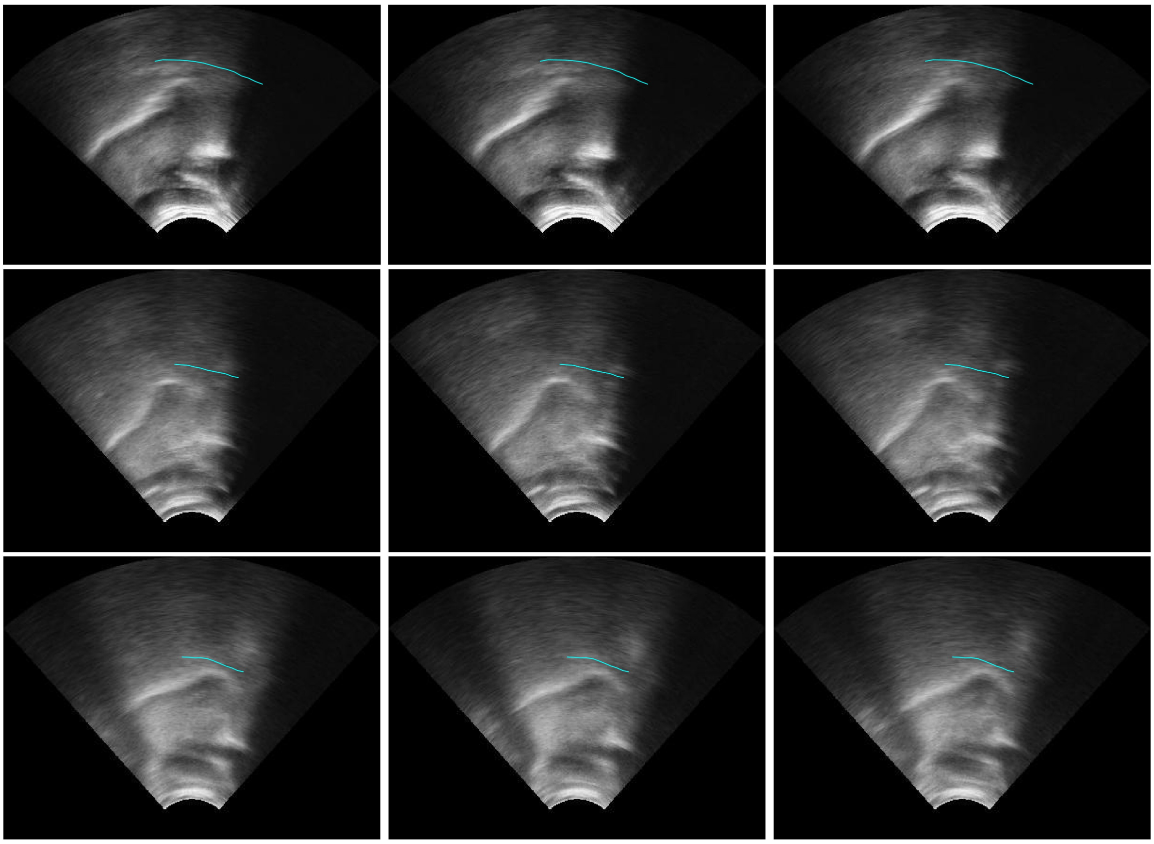


Figure 2: ***Mean tongue posture at rhotic targets**. Mean image calculated from target frame in three repetitions of three words. L-to-R: back vowel contexts [oː-ʊ-ɔ], low vowel contexts [æ-ɐː-ɐ], front vowel contexts [iː-ɪ-e]. Top row: Participant W003; Middle row: W006; Bottom row: W007. Blue line: palate trace. (Front of mouth: right side of image).*

3. Results

Two lingual gestures were observed in all rhotics produced by all speakers in the experimental corpus: dorsal retraction toward a mid- to low-pharyngeal target, and coronal raising toward a target in the alveolar-palatal region. For all speakers, the two gestures were largely synchronous; precise intergestural timings have not yet been quantified. Participants differed in (i) the place of articulation of the coronal gesture, (ii) the tongue shape at rhotic target, and (iii), the degree and type of influence of vowel context on target rhotic posture.

Mean midsagittal images capturing mid-consonantal tongue postures for word-initial /ɹ/ produced by each participant in high front, low, and back vowels contexts are illustrated in Figures 2 and 3. Individual differences in coronal articulation can be observed: more retracted toward a palatal target for Speakers W003, W006, W009 and W014, and more anterior for W007, who uses a post-alveolar coronal gesture in all contexts. Speaker W010 articulates /ɹ/ with an alveolar coronal gesture in back vowel contexts, a more retracted palatal constriction in front vowel contexts, and a more distributed laminar coronal gesture intermediate between these two places in low vowel contexts.

Rhotics produced by all six speakers were characterized by a ‘saddle’ – some degree of concavity in the mid-lingual region – in at least some vowel contexts. For Speakers W003, W006, W007 and W009, all rhotics were produced with a mid-lingual saddle, regardless of vowel context. Speaker W014 articulates /ɹ/ with a more bunched tongue posture showing minimal dorsal concavity. Speaker W010 produces rhotics before low vowels (*‘rap’*–*‘raft’*–*‘rough’*) with a globally convex tongue shape, but other realizations show saddles at different parts of the tongue, depending on context: the concavity appears at a similar mid-lingual location to other speakers before front *‘reef’*–*‘rip’*–*‘ref’*, but at a more anterior part of the tongue before back vowels *‘raw’*–*‘rook’*–*‘ref’* (Fig. 3, middle row).

4. Discussion

Consistent lingual postures during rhotic production were observed in all vowel contexts by five of the six speakers in our study – tongue shapes that broadly correspond to the ‘tip-down’ configuration classified as ‘Type 4’ by Delattre and Freeman [7]. The predominance of this tongue shape is consistent with previous findings for NZE [20] and North American Englishes [11], in which ‘bunched’ variants were also the most common rhotic allophones. In contrast to these previous studies, none of the AusE speakers consistently and exclusively produced ‘tip-up’ variants; however, six speakers is too a small sample from which to draw further conclusions at this stage.

Furthermore, closer inspection of even these initial data suggests that categorical classification of tongue postures into binary taxonomies misrepresents the complexity of articulation involved. Rhotics produced by W009 in non-front vowel contexts, for example, demonstrate some evidence of retroflexion towards a palatal target (Fig. 3, top row), which is not well

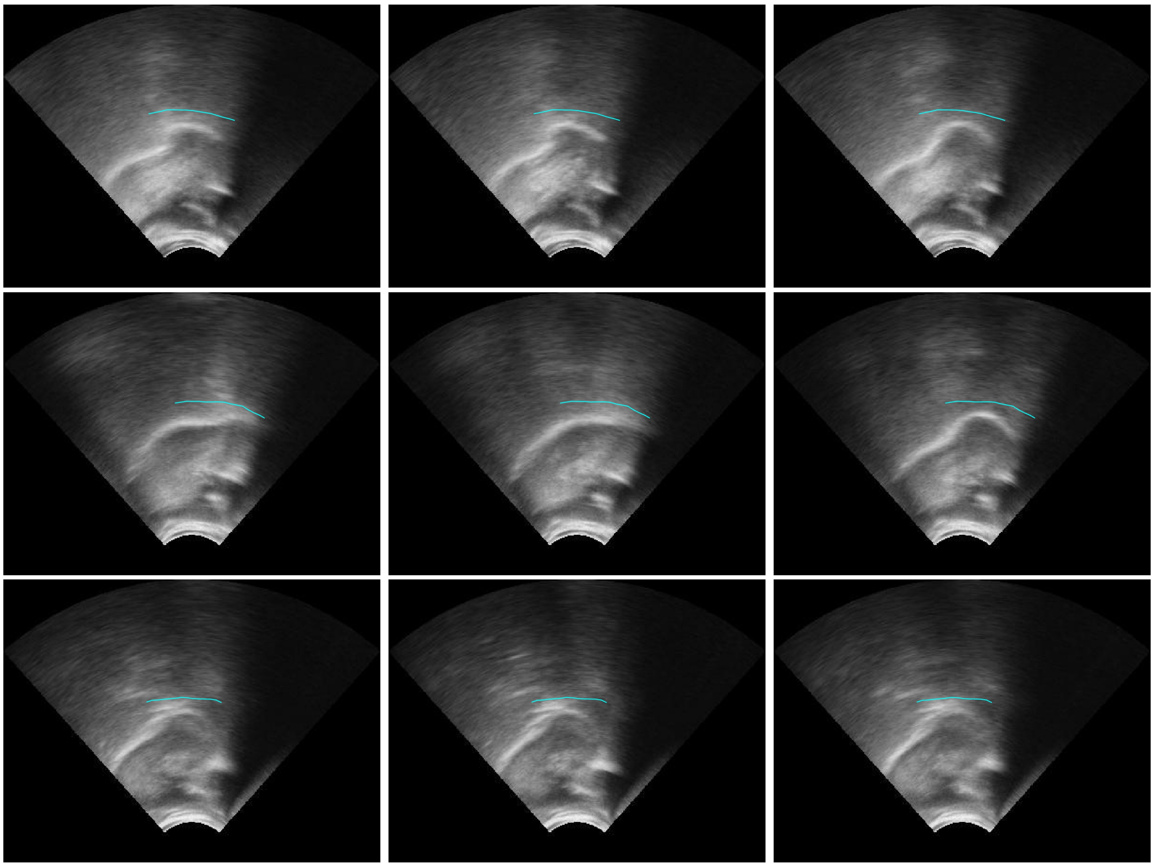


Figure 3: ***Mean tongue posture at rhotic targets***. *Mean image calculated from target frame in three repetitions of three words. L-to-R: back vowel contexts [oː-ʊ-ɔ], low vowel contexts [æ-ɐː-ɐ], front vowel contexts [iː-ɪ-e]. Top row: Participant W009; Middle row: W010; Bottom row: W014. Blue line: palate trace. (Front of mouth: right side of image).*

characterized as ‘bunched’, nor as either ‘tip-up’ or ‘tip-down’. More systematic investigation of tongue shaping in these and other speakers of Australian English is required to better represent articulation in the midsagittal plane, and how this relates to goals of production [28, 29, 30]. Key to this will be dynamic analysis of tongue shaping, including quantification of intergestural timing and examination of how the coordination of lingual gestures shapes the tongue over time in different vowel contexts.

4.1. Methodological considerations

In this study, data were acquired using the AAA system [25] and exported for inspection and analysis using a custom Matlab-based tool (Fig. 1). This allowed for greater flexibility in analysis, but requires robust synchronization of exported audio and video data. While this is usually feasible, platform-independent methods for audio-video synchronization during and after ultrasound recordings are important for validation and maximal flexibility in data analysis and interpretation (e.g. [31]).

The carrier phrase *‘It’s a ___’* worked well for elicitation of onset /ɹ/ in an environment where key lingual gestures could be observed before a range of vowels. It was important to monitor speakers during data collection to ensure that they produced the sentences with a pre-rhotic schwa, as one speaker (W007) produced a diphthongized [æɪ] article in some trials. While this is unlikely to have influenced target tongue postures significantly, it will have affected gestural timing and coarticulation.

The aluminium headset [26] was effective in stabilizing the ultrasound probe, but partly due to the weight (∼0.8 Kg), most speakers experienced some discomfort over the duration of a 40 minute experiment. More ergonomic options (e.g. [32]) may facilitate improved user comfort. Some displacement/rotation of the probe may have affected consistency of the imaging plane and therefore reliability of the data for some speakers, and inconsistency in the relative location of hyoid and mandible shadows across speakers makes direct comparison of place of articulation difficult. No bite plate was used to calibrate ultrasound data with anatomical landmarks [33], so images in Figures 2–3 should be interpreted with caution.

5. Conclusions

This study is a first step towards the systematic investigation of Australian English /ɹ/ production using instrumental methods. Initial investigation of lingual articulation in six speakers reveals two coordinated lingual gestures: a mid pharyngeal tongue body gesture, and a coronal gesture realized at a speaker-specific place of articulation. Consistent tongue shaping was observed across vowel contexts for five speakers, in a posture characterized by some degree of mid-lingual concavity. These data further demonstrate the utility of midsagittal ultrasound tongue imaging as a method for characterizing general goals of rhotic approximants, and for revealing individual speaker variation.

6. Acknowledgements

This project was supported by Macquarie University Research Acceleration Scheme (MQRAS0001053) "Visualising complex speech sounds: a study of rhotic consonants – evidence from ultrasound of the tongue". We would like to thank Khan Milligan who assisted with data collection.

7. References

[1] R. M. Dalston, "Acoustic characteristics of English /w, r, l/ spoken correctly by young children and adults," *JASA*, vol. 57, no. 2, pp. 462–469, 1975.

[2] O. Fujimura and D. Erickson, "Acoustic phonetics," in *The Handbook of Phonetic Sciences*, W. Hardcastle and J. Laver, Eds. Oxford: Blackwell, 1997, pp. 65–115.

[3] B. Heselwood and L. Plug, "The Role of F2 and F3 in the Perception of Rhoticity: Evidence from Listening Experiments," in *Proc. ICPhS*, 2011, pp. 867–870.

[4] C. Y. Espy-Wilson, S. E. Boyce, M. Jackson, S. Narayanan, and A. Alwan, "Acoustic modeling of American English /r/," *JASA*, vol. 108, no. 1, pp. 343–356, 2000.

[5] P. J. Howson and P. J. Monahan, "Perceptual motivation for rhotics as a class," *Speech Communication*, vol. 115, pp. 15–28, 2019.

[6] K. Sebregts, R. van Hout, and H. Van de Velde, "Sociophonetics and rhotics," in *The Routledge Handbook of Sociophonetics*. Routledge, 2023, pp. 195–213.

[7] P. Delattre and D. C. Freeman, "A dialect study of American r's by x-ray motion picture," *Linguistics*, vol. 44, pp. 29–68, 1968.

[8] M. K. Tiede, S. E. Boyce, C. K. Holland, and K. A. Choe, "A new taxonomy of American English /r/ using MRI and ultrasound," *JASA*, vol. 115, no. 5, pp. 2633–2634, 2004.

[9] X. Zhou, C. Y. Espy-Wilson, S. Boyce, M. Tiede, C. Holland, and A. Choe, "A magnetic resonance imaging-based articulatory and acoustic study of 'retroflex' and 'bunched' American English /r/," *JASA*, vol. 123, no. 6, pp. 4466–4481, 2008.

[10] E. Lawson, J. M. Scobbie, and J. Stuart-Smith, "The social stratification of tongue shape for postvocalic /r/ in Scottish English," *J. Sociolinguistics*, vol. 15, no. 2, pp. 256–268, 2011.

[11] J. Mielke, A. Baker, and D. Archangeli, "Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /r/," *Language*, vol. 92, no. 1, pp. 101–140, 2016.

[12] A. Alwan, S. Narayanan, and K. Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics," *JASA*, vol. 101, no. 2, pp. 1078–1089, 1997.

[13] M. Proctor and R. Walker, "Articulatory bases of English liquids," in *The Sonority Controversy*, S. Parker, Ed. Berlin: De Gruyter, 2012, vol. 18, pp. 285–312.

[14] B. Gick, "Articulatory correlates of ambisyllabicity in English glides and liquids," in *Papers in laboratory phonology VI: Phonetic interpretation*, J. Local, R. Ogden, and R. A. M. Temple, Eds. Cambridge: CUP, 2003, pp. 222–236.

[15] K. Iskarous, "The Articulation of the Palatal Gesture in American English [r]," in *Proc. 7th Intl. Seminar on Speech Production*, Ubatuba, 2006.

[16] B. Gick, K. Iskarous, D. H. Whalen, and L. M. Goldstein, "Constraints on variations in the production of English /r/," in *6th Intl. Seminar on Speech Production*, Sydney, 2003, pp. 73–78.

[17] B. Gick, F. Campbell, S. Oh, and L. Tamburri-Watt, "Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids," *J. Phon*, vol. 34, no. 1, pp. 49–72, 2006.

[18] M. Proctor, R. Walker, C. Smith, T. Szalay, S. Narayanan, and L. Goldstein, "Articulatory characterization of English liquid-final rimes," *J. Phon*, vol. 77, p. 100921, 2019.

[19] F. Cox and S. Palethorpe, "Australian English," *JIPA*, vol. 37, no. 03, pp. 341–350, 2007.

[20] M. Heyne, X. Wang, D. Derrick, K. Dorreen, and K. Watson, "The articulation of /ɹ/ in New Zealand English," *JIPA*, vol. 50, no. 3, pp. 366–388, 2020.

[21] F. Cox, L. Ratko, J.-H. Kim, J. Penney, and M. Proctor, "Investigating rhotic production by Australian English speakers using ultrasound imaging," in *Australian Linguistic Society Annual Conf.*, University of Sydney, 29 Nov - 01 Dec. 2023.

[22] K. Crowe and S. McLeod, "Children's English consonant acquisition in the United States: A review," *American J. Speech-Language Pathology*, vol. 29, no. 4, pp. 2155–2169, 2020.

[23] S. Lin and K. Demuth, "Children's acquisition of English onset and coda /l/: Articulatory evidence," *JSLHR*, vol. 58, no. 1, pp. 13–27, 2015.

[24] P. J. Howson and M. A. Redford, "The acquisition of articulatory timing for liquids: Evidence from child and adult speech," *JSLHR*, vol. 64, no. 3, pp. 734–753, 2021.

[25] Articulate Instruments, *Articulate Assistant Advanced User Guide: Version 220.2.0.*

[26] J. M. Scobbie, A. A. Wrench, and M. van der Linden, "Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement," in *Proc. 8th Intl. Seminar on Speech Production*, 2008, p. 373–376.

[27] M. A. Epstein and M. Stone, "The tongue stops here: Ultrasound imaging of the palate," *JASA*, vol. 118, no. 4, pp. 2128–2131, 2005.

[28] I. Stavness, B. Gick, D. Derrick, and S. Fels, "Biomechanical modeling of English /r/ variants," *JASA*, vol. 131, no. 5, pp. EL355–EL360, 2012.

[29] S. Stolar and B. Gick, "An index for quantifying tongue curvature," *Canadian Acoustics*, vol. 41, no. 1, 2013.

[30] K. M. Dawson, M. K. Tiede, and D. Whalen, "Methods for quantifying tongue shape and complexity using ultrasound imaging," *Clinical Linguistics & Phonetics*, vol. 30, no. 3-5, pp. 328–344, 2016.

[31] A. Eshky, J. Cleland, M. S. Ribeiro, E. Sugden, K. Richmond, and S. Renalds, "Automatic audiovisual synchronisation for ultrasound tongue imaging," *Speech Communication*, vol. 132, pp. 83–85, 2021.

[32] L. Spreafico, M. Pucher, and A. Matosova, "Ultrafit: A speaker-friendly headset for ultrasound recordings in speech science," in *Proc. Interspeech*, 2018, pp. 1917–1520.

[33] L. Ménard, J. Aubin, M. Thibeault, and G. Richard, "Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model," *Folia Phoniatrica et Logopaedica*, vol. 64, no. 2, pp. 64–72, 2012.

Give a Little Whistle: A Neglected Characteristic of Australian English Productions of /s/

Elise Tobin[1], *Joshua Penney*[1], *Hannah White*[1] *and Felicity Cox*[1]

[1]Department of Linguistics, Macquarie University, Sydney, Australia

elise.tobin@hdr.mq.edu.au, joshua.penney@mq.edu.au, hannah.white@mq.edu.au, felicity.cox@mq.edu.au

Abstract

This acoustic-phonetic study examined the presence of whistle in Australian English (AusE) adult and child productions of /s/. We report on whistle identification based on visual and auditory analysis, and we consider how the presence of whistle can impact kurtosis (M_4), a lesser-studied spectral measure known to correlate with whistle presence. Results of linear mixed models indicated that tokens with whistle characteristics had higher kurtosis (M_4) values compared to tokens without whistle, with no differences found for age or gender groups. Whistled tokens in rounded vowel contexts had lower M_4 values compared to whistled tokens in non-rounded vowel contexts.

Index Terms: fricatives, kurtosis, whistle, Australian English, vowel contexts, gender, children's speech

1. Introduction

Whistled fricatives are characterised by the simultaneous occurrence of a whistling component along with frication during production. Fricatives are characterised by a continuous noise source resulting from constriction in the vocal tract [1]. The fricatives /s, z, ʃ/ can be produced with whistle [1, 2, 3], although this is perhaps more common for the voiceless alveolar fricative /s/, which has phonemic variations in languages like Xitsonga and Changana [4, 5] and non-phonemic variations in language like English and French [6, 7]. /s/ is produced with a narrow constriction, formed by raising the tongue blade to the alveolar ridge, resulting in turbulent airflow (frication) which is channelled through the deep, narrow groove on the tongue dorsum and amplified by the lower teeth obstacle [8, 9, 10]. The frication noise is characterised by aperiodic energy concentrated at high frequencies above 4kHz [9]. Fricative spectral shapes are based on the shape and size of the cavity in front of the constriction [11]. Alveolar fricatives have long front cavities (in comparison to labiodental or dental fricatives), which result in clearly defined spectral shapes [11].

There may be subtle differences in articulatory configurations in productions of non-whistled and whistled /s/. Articulatory configurations for productions of English /s/ have been shown to vary between speakers and in different vowel contexts [3]. Although we do not know the articulatory configurations of non-phonemic whistled fricatives, [12] suggests that a raised tongue tip can produce "edge tone" whistled fricatives. The "edge tone" model is one of the proposed mechanisms for how whistles are produced when an unstable jet of air strikes an obstacle [1, 12]. For whistled fricatives, the tongue constriction channels the jet of air, and the teeth are the obstacle (or edge) [1, 12]. The turbulent air oscillates around the teeth and couples with the resonances of the cavity between the tongue constriction and teeth, resulting in a whistled fricative [1, 4, 12, 13]. Small changes in articulatory configuration will affect the front cavity resonances [6, 9], which can be measured with spectral moments analysis.

Spectral moments analysis allows us to examine the distribution of energy in the fricative [14]. Four spectral moments are calculated from the distribution; the mean energy (M_1), also referred to as centre of gravity (COG), variance (M_2), skewness (M_3) and kurtosis (M_4); only M_4 shall be analysed here. Kurtosis (M_4) is a measure of the peakiness of the spectrum (i.e., the distribution) [15]. A clearly defined spectrum with higher peaks will have a positive kurtosis (i.e., higher M_4 values), while a less defined spectrum without clear peaks will have a negative kurtosis (i.e., lower M_4 values) [11]. Kurtosis is rarely reported in analyses of spectral moments of fricatives, with many studies providing data on a limited number of spectral moments, or in some cases only the first spectral moment, COG [7, 15]. However, kurtosis may be a useful metric for identifying whistle presence in fricatives [16].

Although whistling is thought to be relatively common in fricative production [13], non-phonemic whistled fricatives are rarely reported in fricative analysis studies. One recent study conducted an exploratory analysis on the proportion of whistled /s/ and /z/ in whispered and voiced speech produced by adult female Portuguese speakers [2]. The corpus consisted of sustained fricatives and fricatives in onset, medial and coda position. Whistled tokens were identified by the presence of a spectral peak between 9 and 13kHz. [2] reported a high proportion of whistled tokens in their dataset – 402 whistled /s/ tokens (219 voiced, 183 whispered) compared to 166 non-whistled /s/ tokens (69 voiced, 97 whispered), and 154 whistled /z/ tokens compared to 113 non-whistled /z/. Another recent study reported on the ultra-high frequency whistled [s͎], characterised by a COG above 10kHz, in southern Chilean Spanish [16]. [16] reported that non-phonemic whistled [s͎] comprised 9% of their corpus, with female speakers more likely to produce this phone compared to male speakers. Lower socioeconomic status was associated with increased use of whistled [s͎], indicative of a sociophonetic function [16].

The spectra of whistled fricatives (phonemic and non-phonemic) are described to have high-amplitude and narrow-bandwidth peaks [1, 5]. However, extracting and interpreting spectral slices for individual whistled fricative tokens in a corpus can be a time-consuming process, and it may not be necessary to inspect spectra to confirm the auditory percept of a whistled fricative. For instance, the Changana phonemic whistled fricative /sv/ is described to have a dark horizontal band below 5000Hz on the spectrogram [5]. It may also be possible to identify non-phonemic whistled /s/ based on audible whistling and spectrographic cues such as this. Kurtosis has been suggested as a key measure for distinguishing between /s/ and the ultra-high frequency whistled [ṣ] in Chilean Spanish, with higher values observed for the whistled variant [16]. Kurtosis values may also reflect differences in tongue posture in /s/ productions [17]. In [17], English /s/ was reported to have higher M_4 values compared to Japanese /s/, with the difference proposed to stem from slight variations in articulatory configurations between English /s/ (more apical) and Japanese /s/ (more laminal) productions.

[6, 7] suggest that non-phonemic whistled fricatives occur more frequently in rounded vowel contexts. In [6], a single adult French-speaking male participant produced a whistled /s/ in the context of /usu/, and the spectra was observed to have a high-amplitude, narrow-bandwidth peak at 8.2kHz. [7] reported that kurtosis varied across speakers, and it was suggested that this variance may be partially due to whistled fricatives occurring in rounded contexts. Kurtosis was reported to be higher in whistled /s/ tokens produced by a male adolescent participant in their study [7]. The number of whistled fricatives produced by this single participant was not reported, and it is unclear if other participants also produced whistled fricatives.

The main research question of this paper is whether the presence of non-phonemic whistle in the English voiceless alveolar fricative /s/ is associated with high kurtosis (M_4) values. It was hypothesised that kurtosis (M_4) would be higher in tokens with visual and auditory features of whistle presence, compared to tokens without these features. Two studies have observed non-phonemic whistled /s/ tokens produced by an adolescent [7] and an adult speaker [6]. Gender differences have been observed for some spectral moments (M_1-M_3) of /s/ produced by Australian-English speaking children [18]. In the present study we analysed tokens produced by adult and child speakers to determine whether non-phonemic whistled /s/ differs across age groups and whether females produce more whistled /s/ tokens compared to male speakers, as observed for the ultra-high frequency [ṣ] in Chilean Spanish [16]. The second exploratory research question of this paper is whether M_4 values in whistle tokens are moderated by vowel context. As previous observations of non-phonemic whistled /s/ have noted the impact of vowel rounding [6, 7], we predict that M_4 values of whistle tokens would be higher in the more rounded vowel context (/ɔ/), compared to the unrounded contexts (/ɪ, æ, ɐ/). Kurtosis may reflect variability in fricative production, highlighting the importance of this spectral measure.

2. Method

2.1. Participants

Participants were Australian English-speaking adults and children. A total of 36 adults between the ages of 18-40 (M_{age}= 24; F = 25, M = 10, Gender Unspecified = 1) and 19 children between the ages of 4 years – 5 years;11 months (M_{age} = 57 months; F = 10, M = 9) participated. Participants were recruited in Sydney and were reimbursed with either course credit (where applicable) or a $30 voucher (per session). Adult participants were recruited through Macquarie University undergraduate units and via Macquarie University channels. Child participants were recruited from local childcare centres and via advertisements. All adult participants had completed their schooling in Australia and at least one parent/caregiver of child participants had completed all schooling in Australia. Participants had no history of speech disorders or hearing impairment. Four adults were excluded due to incomplete questionnaires, history of speech intervention, or non-binary gender identification. Child participant's hearing was screened using the Sound Scouts iPad application [19] and only those who obtained a 'Pass' result were included in this study (4 children were excluded on this basis). 15 child participants were included in this study (M_{age}= 57.5 months; F = 7, M = 8). This production study was conducted within the context of a larger PhD project.

2.2. Stimuli

The target items consisted of four CVC words, with the voiceless alveolar fricative /s/ in onset position, followed by a short monophthong: (two front unrounded vowels /ɪ, æ/; one central unrounded vowel /ɐ/; one back rounded vowel /ɔ/) (*sip*, *sap*, *suck* and *sock*). Although the target items were not specifically selected to examine vowel rounding effects, /ɔ/ (*sock*) is likely to be more rounded in AusE than the other vowels [20], providing an opportunity to explore this feature. Four additional CVC target words with the voiceless alveolar stop /t/ in onset position were also elicited; these are not analysed in the current study. The target words were produced in utterance final position in the carrier phrase "A happy X". In addition to the target items, 7 non-target CVC words were used for practice trials.

2.3. Procedure

The study was conducted in a child-friendly, acoustically attenuated lab in the presence of the first author, with caregivers of child participants able to observe from a connected room. Participants were seated at a desk in front of a computer screen on which the experiment was presented, using a customised Praat script controlled by the experimenter. Recordings were made with a Neumann TL103 condenser microphone at a sampling rate of 44.1kHz. The microphone was mounted on an articulated microphone boom and positioned approximately 20cm from the participant's mouth. Prior to the production task, the experimenter showed participants images of each target word while modelling the target word, and the participant was asked to repeat the word. Participants were informed that they would be prompted with the correct word if required during the task. Auditory instructions and stimuli prompts to be used in the experiment were produced by a 30-year-old female Australian English-speaker in a child-directed manner and recorded in an acoustically attenuated lab at a sampling rate of 44.1kHz. Stimuli prompts were the natural /s/ productions of this speaker.

The visual stimulus for each trial consisted of a sad cartoon panda on the left side of the screen and the image of the target item on the right side of the screen. A short introduction to the experiment familiarised participants with the carrier phrase "A happy X". An audio recording informed the participant that they could help make the panda smile by saying the name of the objects next to the panda. On each practice trial, the participant saw the visual stimulus and heard an auditory prompt, e.g., "Make the panda happy! Say, a happy pig", and participants would repeat the phrase "A happy X". On each test trial, the visual stimulus was paired with the auditory prompt "Make the panda happy!", and participants produced the carrier phrase and target word. If the participant could not remember the target word or produced the target word in isolation, an auditory stimulus prompt would be played, e.g., "Say, a happy sip". Once the participant produced the carrier phrase and target word, the image of the sad panda would change to an image of a happy panda at the end of each trial. Adult and child participants completed the same task, differing only in the number of trials. We retrospectively analysed the stimuli prompts, which revealed a slight whistle in the *sip* context. Of the total 403 *sip* tokens, 19 tokens were produced after hearing the prompt; of these, 3 tokens were produced with whistle, all by children.

Child participants completed 6 blocks of trials which contained the 8 target words in a pseudo-randomised order, for a total of 48 tokens (8 words x 6 repetitions) per child participant; 24 of these were /s/-target words. Adult participants completed 10 blocks of trials, for a total of 80 tokens (8 words x 10 repetitions) per adult participant; 40 of these were /s/-target words.

2.4. Analysis

Audio files and corresponding text files were processed using the BAS web services pipeline to generate phoneme aligned TextGrids [21]. Using Praat [22], boundaries for the /s/ phoneme were individually inspected and corrected when these did not align with the phoneme onset and offset. The /s/ onset boundary was placed at the start of strong frication and the /s/ offset boundary was placed at the end of strong frication on the spectrogram. The spectrogram view range was increased to 16kHz when coding for whistle presence to ensure that this characteristic was identified consistently across adult and child tokens.

Each token was inspected to determine whether whistle was present or absent. Whistle presence was indicated by an audible whistle during part of the fricative and the presence of a strong band of energy in the spectrogram. Measurements obtained included measures of duration, intensity, spectral peak and the four spectral moments (M_1-M_4). Spectral measures were obtained by processing audio files and corresponding TextGrids with a customised Praat script [23]. This script was originally adapted from DiCanio [24] with adjustments for some of the formulas, including window spacing and bin calculations, and subsequently updated to include the revised calculation methods in version 4.0 of the DiCanio script [25]. Potential effects of co-articulatory voicing on spectral measures [26] were reduced by high-pass filtering target sounds at 300Hz within the Praat script. Spectral measures were calculated using time averaging of windows, as recommended by Shadle [13]. Each sound had a trim of 25% for both onset and offset, i.e., the central 50% of each sound domain was used for spectral measure calculations. Ten rectangular windows with a 5ms size were spaced evenly across this sound domain. The focus here is on kurtosis (M_4), as it may be sensitive to the presence of whistle [6, 7, 16].

There were 35 tokens excluded from analyses (e.g., for noise, mispronunciations, pauses, etc.). The analysis was conducted on a total of 1605 tokens (Adults = 1272, Children = 333).

2.5. Statistical Analysis

The analysis reported here assesses the effect of speaker age, gender, whistle presence and vowel context on the dependent variable, kurtosis (M_4). Linear Mixed Effects Regression (LMER) analysis of kurtosis was conducted in R Studio [27] using the lme4 package [28]. Degrees of freedom for t-tests were estimated with Satterthwaite's method using the lmerTest package [29]. The fixed effects in the model were group (i.e., Adults vs Children), gender (i.e., Female vs Male), whistle presence (i.e., Absent vs Present in token), vowel context (i.e., /æ/ vs /ɪ, ɐ, ɔ/; *sap* vs *sip*, *suck, sock*) and fricative duration. The model included an interaction term for group and gender, and an interaction term for whistle presence and vowel context. The random effects structure included intercepts for participants.

3. Results

A total of 180 tokens were identified to have whistle present (Adults = 116, Children = 64). Whistle tokens were produced by 62% of the participants (19 Adults, 10 Children). Table 1 provides the total number of tokens with and without whistle in the analysis, by group and gender. The number of tokens with whistle presence was roughly equivalent across vowel contexts (/ɪ/ = 11%, /æ/ = 12%, /ɐ/ = 10%, /ɔ/ = 12%).

Table 1. *Tokens with and without whistle*

Group	Gender	Whistle Absent	Whistle Present	% with Whistle
Adults	Female	781	92	11%
Adults	Male	375	24	6%
Children	Female	128	32	20%
Children	Male	141	32	18%
Total =		1425	180	11%

Adult speakers had a mean kurtosis (M_4) value of 3.6 (Female M=3.7, SD=5.7; Male M= 3.6, SD=2.9), and child speakers had a mean M_4 value of 2.98 (Female M=4.61, SD=9.68; Male M=1.48, SD=2.13). Figure 1 provides the kurtosis (M_4) values by whistle presence and vowel context, averaged over group and gender, and the mean M_4 value represented with purple circles. Some tokens that were labelled as 'Whistle Absent' during acoustic coding have relatively high M_4 values between 10 – 20. Although these 'Whistle Absent' tokens did not have an audible whistle or clear bands of energy in the spectrogram, the high M_4 values may indicate that these tokens were produced with a similar mechanism as those produced with audible/visible whistle characteristics.

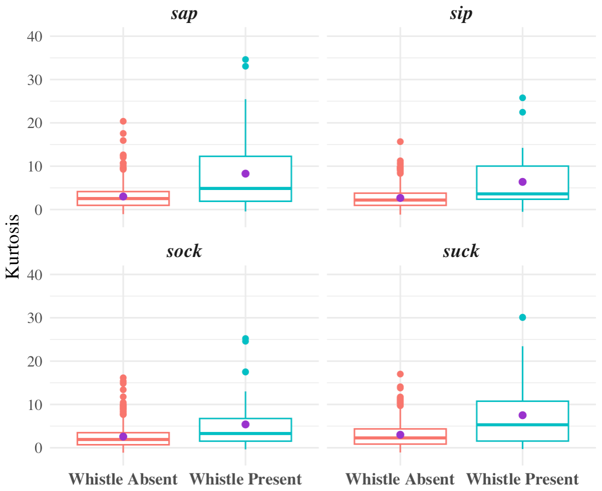


Figure 1: *Kurtosis (M_4) by whistle presence & vowel context.*

The analysis of speaker age, gender, whistle presence and vowel context on kurtosis (M_4) found a significant interaction between whistle presence and vowel context (β = -4.1, SE= 0.955, t(1554)= -4.27, p < 0.001). *Post hoc* comparisons were run using the emmeans package [30] in R, with Tukey HSD corrections to *p*-values for multiple comparisons. Our first research question asked whether whistle in /s/ tokens would predict higher M_4 values, and research question 2 asked whether this was moderated by vowel rounding. In support of our prediction for research question 1, pairwise comparisons indicated that M_4 values were significantly higher in tokens with whistle present, compared to tokens without whistle present (all vowel contexts $p < 0.001$). In contrast to the prediction for research question 2, whistles in *sock* tokens have a significantly lower M_4 compared to the M_4 values of whistle tokens in other vowel contexts ($p < 0.05$). This interaction may indicate that in the /ɔ/ vowel context, there are other (perhaps more prominent) cues to whistle, such as the other spectral moments (M_1, M_2, M_3) or other acoustic measures. There were no other significant simple effects or interactions, indicating there is no evidence that M_4 values differ based on group or gender for this dataset.

Two example spectrograms are provided here, both with a view range of 16kHz, to illustrate the range of whistle characteristics. Figure 2 is an example of a female adult's token with whistle present and a strong band of energy visible around 8kHz. This token had an audible whistle during the fricative and a high M_4 value of 85.

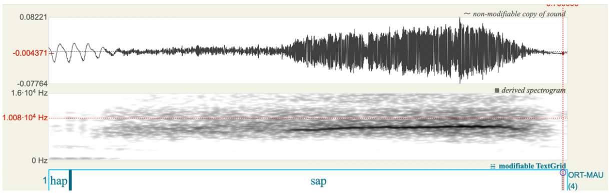


Figure 2: *Whistle presence in female adult's token.*

Figure 3 is an example of a female child's token with whistle present and a strong band of energy visible between 8-10kHz. This token had an audible whistle during the fricative and a M_4 value of 21.

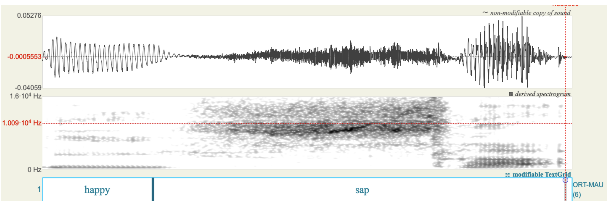


Figure 3: *Whistle presence in female child's token.*

4. Discussion

This study aimed to investigate whether the presence of whistle in /s/ productions is associated with higher kurtosis (M_4). Consistent with previous observations of non-phonemic whistled /s/ [6, 7], this study has provided evidence that high M_4 values may be indicative of /s/ produced with whistle characteristics. The LMER analysis indicated that whistled /s/ tokens had higher M_4 values, compared to non-whistled /s/ tokens, across all vowel contexts. Some non-whistled /s/ tokens were observed to have relatively high M_4 values, which may indicate that these tokens were produced with a similar articulatory configuration to whistled /s/ tokens. Some studies have observed whistled /s/ in rounded contexts [6, 7], while this study has observed whistled /s/ in both unrounded vowel contexts (i.e., /ɪ, æ, ɐ/) and rounded contexts (i.e., /ɔ/). There was a significant interaction of M_4 values and vowel context; contrary to our expectations, M_4 values in whistled tokens in the *sock* context were lower compared to M_4 values of whistle tokens in the *suck*, *sap* and *sip* contexts. This is despite high kurtosis being previously reported, and confirmed here, as being associated with whistled fricatives, and suggestions that whistling may be more common in the context of rounded vowels [6, 7]. It is possible that the lower M_4 values seen here in the /ɔ/ context for whistle tokens may reflect the backness of the vowel, which is produced with a more retracted tongue position [20] compared to the front and central vowels (/ɪ, æ, ɐ/). Future research should examine whistling in fricatives in a greater range of vowel contexts, both rounded and unrounded.

The total proportion of whistled /s/ tokens in this study was comparable to the proportion of whistled [ʂ̩] in Chilean Spanish [16]. Whistled /s/ was relatively common in both age groups, with children observed to have a higher proportion of whistled /s/ compared to adults. Female adults were observed to have a higher proportion of whistled /s/ compared to male adults, as was also reported for Chilean Spanish speakers [16]. However, there were insufficient tokens in this data set for a logistic regression analysis to assess the significance of the observed age and gender differences. Future research should examine whether these age and gender differences are observed in larger corpora of Australian English and in spontaneous, conversational speech.

There was a lower proportion of whistled /s/ observed in this study compared to the proportion of whistled /s/ reported to occur in whispered and voiced speech of female Portuguese speakers [2]. This may reflect differences in elicitation methods, as this study did not elicit whispered speech, and /s/ was only elicited in onset position, while [2] elicited /s/ in onset, medial and final position in whispered and voiced speech, as well as eliciting sustained /s/ tokens. Further research is required to determine whether whistled /s/ occurs in medial and final position to the same or greater extent than /s/ in onset position. The higher proportion of whistled /s/ for Portuguese speakers may also reflect a subtle difference in the production of /s/ by English and Portuguese speakers.

Kurtosis is infrequently reported in studies on fricatives, perhaps due to the perception that kurtosis does not capture sociophonetic features [15]. However, this may be a misperception, as kurtosis is a key measure for identifying non-phonemic whistled [ʂ̩] in Chilean Spanish, which is associated with female speech [16]. This study has demonstrated that kurtosis is also an important measure for identifying non-phonemic whistled fricatives in Australian English. Further research is required to determine whether sociophonetic factors are associated with English non-phonemic whistled /s/. The neglect of kurtosis may need to be reconsidered when describing typical fricative production by children and adults. Whistled /s/ was relatively common in both age groups and kurtosis appears sensitive to this phenomenon, which suggests it may be a useful metric for capturing variability in fricative production. Analysing kurtosis may be a more time-efficient method for identifying whistle presence in fricatives, compared to extracting and interpreting individual spectral slices. Future studies of fricative production would benefit from identifying whistled tokens and including kurtosis in analyses of spectral moments, as it may improve our knowledge of sociophonetic variation in English.

5. Acknowledgements

We thank Rebecca Holt for providing the speech stimuli used in this study, and Titia Benders for assisting with the development of the customised Praat script used in the experiment. Thanks to Macquarie University Child Language Lab for helpful discussions on methodology and Macquarie University Phonetics Lab for feedback on data analysis, in particular Sean Boylan, who provided the customised spectral measures Praat script and valuable advice on spectral moments analysis. This research was supported by a government funded Research Training Program Scholarship #202113239 to the first author, and an ARC Future Fellowship Grant FT180100462 and a Discovery Project Grant DP190102164 to the fourth author.

6. References

[1] Shadle, C. H., "The Aerodynamics of Speech", in W. J. Hardcastle, J. Laver, and F. E. Gibbon [Eds], The Handback of Phonetic Sciences, 39-80, Blackwell Publishing Ltd, 2010.

[2] Jesus, L. M. T., Castilho, S., Ferreira, A. J. S., and Costa, M. C., "Attributes Associated with Consonantal Place and Voicing in Whispered Speech", in 13th International Seminar of Speech Production, 2024.

[3] Proctor, M., Shadle, C. H., and Iskarous, K., "An MRI study of vocalic context effects and lip rounding in the production of English sibilants", in Proceedings of the 11th Australian International Conference on Speech Science & Technology: Australian Speech Science & Technology Association Inc., 2006.

[4] Lee-Kim, S., Kawahara, S., and Lee, S. J., "The 'Whistled' Fricative in Xitsonga: Its Articulation and Acoustics", Phonetica, 71(1): 50–81, 2014, doi: 10.1159/000362672.

[5] Shosted, R. K., "Articulatory and acoustic characteristics of whistled fricatives in Changana", in Selected Proceedings of the 40th Annual Conference on African Linguistics: African Languages and Linguistics Today, 2011.

[6] Shadle, C. H. and Scully, C., "An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences", Journal of Phonetics, 23(1-2):53–66, 1995. https://doi.org/10.1016/S0095-4470(95)80032-8.

[7] Koenig, L. L., Shadle, C. H., Preston, J. L., and Mooshammer, C. R., "Toward improved spectral measures of /s/: Results from adolescents", J Speech Lang Hear Res, 56(4):1175–1189, 2013. doi: 10.1044/1092-4388(2012/12-0038).

[8] Shadle, C. H., "The Acoustics of Fricative Consonants", Doctor of Philosophy, Massachusetts Institute of Technology, Cambridge, 1985.

[9] Tabain, M., "Variability in Fricative Production and Spectra: Implications for the Hyper- and Hypo-and Quantal Theories of Speech Production", Lang Speech, 44(1):57–93, 2001. doi: 10.1177/00238309010440010301.

[10] Shadle, C. H., Proctor M. I., Iskarous, K., "An MRI Study of the Effect of Vowel Context on English Fricatives", The Journal of the Acoustical Society of America, 123(5):3735, 2008. https://doi.org/10.1121/1.2935246

[11] Jongman, A., Wayland, R., and Wong, S., "Acoustic characteristics of English fricatives", The Journal of the Acoustical Society of America, 108(3):1252-1263, 2000. https://doi.org/10.1121/1.1288413

[12] Shosted, R. K., "Just Put Your Lips Together and Blow? The Whistled Fricatives of Southern Bantu", UC Berkeley Phonology Lab Annual Reports, 2: 2006. doi: 10.5070/P73P19W08R.

[13] Shadle, C. H., "The Acoustics and Aerodynamics of Fricatives", in A. Cohn, C. Fougeron, and M. K. Huffman [Eds], The Oxford Handbook of Laboratory Phonology, 511-526, Oxford University Press, 2012.

[14] Forrest K., Weismer, G., Milenkovic, P., and Dougall, R. N., "Statistical analysis of word-initial voiceless obstruents: Preliminary data", The Journal of the Acoustical Society of America, 84(1):115-123, 1988. https://doi.org/10.1121/1.396977

[15] Kendall, T., and Fridland, V., "Sociophonetics and Its Methods: Vowels and Sibilants", in Sociophonetics, 40-72, Cambridge University Press, 2021. doi: 10.1017/9781316809709.

[16] Perdomo-Pinto, L., and Sadowsky, S., "The Ultra-High-Frequency Whistled /s/ of Southern Chilean Spanish: Socioeconomic and Gender Stratification of its Spectral Moments and Prevalence", in Proceedings of the 19th International Congress of Phonetic Sciences: International Phonetic Association, 48-52, 2019.

[17] Li, F., Edwards, J., and Beckman, M. E., "Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers", Journal of Phonetics, 37(1):111–124, 2009. doi:10.1016/j.wocn.2008.10.001.

[18] Ford, C., and Tabain, M., "Spectral features of voiceless fricatives produced by Australian English-speaking children", in Proceedings of the 19th International Congress of Phonetic Sciences, International Phonetic Association, 3105-3109, 2019. https://doi.org/10.26181/5f8e7f89cb491

[19] Dillon, H., Mee, C., Moreno, J. C., and Seymour, J., "Hearing tests are just child's play: the sound scouts game for children entering school", International Journal of Audiology, 57(7):529–537, 2018. doi: 10.1080/14992027.2018.1463464.

[20] Blackwood Ximenes, A., Shaw, J. A., and Carignan, C., "A comparison of acoustic and articulatory methods for analyzing vowel differences across dialects: Data from American and Australian English", The Journal of the Acoustical Society of America, 142(1):363–377, 2017. doi: 10.1121/1.4991346.

[21] Schiel, F., Draxler, C., and Harrinton, J., "Phonemic Segmentation and Labelling using the MAUS Technique", in Workshop New Tools and Methods for Very-Large-Scale Phonetics Research, Philadelphia, USA, 2011. doi: https://doi.org/10.5282/ubm/epub.13684.

[22] Boersma, P., and Weenink, D. J. M., "PRAAT, a system for doing phonetics by computer", Glot International, 5, 341-345, 2001. Online: https://www.researchgate.net/publication/208032992_PRAAT_a_system_for_doing_phonetics_by_computer

[23] Boylan, S. P., "/s/ retraction in the /stɹ/ onset in Australian English: is sound change in progress?", Master of Research Thesis, Macquarie University, Sydney, Australia, 2018.

[24] DiCanio, C., "Extract Fricative data from labelled points", 2013. Online: www.buffalo.edu/~cdicanio/scripts/Time_averaging_for_fricatives.praat

[25] DiCanio, C., "Spectral means of fricative spectra script in Praat", 2021. Online: https://www.acsu.buffalo.edu/~cdicanio/scripts/Time_averaging_for_fricatives_4.0.praat

[26] Stevens, M., and Harrington, J., "The phonetic origins of /s/-retraction: Acoustic and perceptual evidence from Australian English", Journal of Phonetics, 58:118–134, 2016. doi: 10.1016/j.wocn.2016.08.003.

[27] R Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, 2023.

[28] Bates, D., Mächler, M., Bolker, B., & Walker, S., "Fitting Linear Mixed-Effects Models Using Lme4," Journal of Statistical Software, 67:1-48, 2015. https://doi.org/10.18637/jss.v067.i01

[29] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B., "lmerTest Package: Tests in Linear Mixed Effects Models," Journal of Statistical Software, 82(13):1-26, 2017. https://doi.org/doi:10.18637/jss.v082.i13

[30] Lenth, R., "emmeans: Estimated Marginal Means, aka Least-Squares Means", R package version 1.10.1, 2024. Online: https://CRAN.R-project.org/package=emmeans.

Vowel Acoustics in Conversational Central Australian Aboriginal English

Yizhou Wang[a], *Carmel O'Shannessy*[b], *Vanessa Davis*[c], *Rikke Bundgaard-Nielsen*[a], *Denise Foster,*[c] *Joshua Roberts*[b]

[a]University of Melbourne, [b]Australian National University, [c]Tangentyere Council

wyz2014tw@gmail.com, Carmel.OShannessy@anu.edu.au, vanessa.davis@tangentyere.org.au, rikkelou@gmail.com, denise.foster@tangentyere.org.au, josh.ian.roberts@gmail.com

Abstract

This paper presents a descriptive study of the acoustic features of the monophthongs in Central Australian Aboriginal English (CAAE, Alice Springs). We extracted vowel tokens from conversational speech of six female speakers who spoke CAAE as one of their primary languages. We analysed the acoustic differences between the vowel categories using a classification analysis and a series of acoustic comparisons based on both duration and spectral information. The results suggest that the phonemic contrasts between all neighbouring monophthongs in CAAE are maintained, and that CAAE vowels display systematic allophonic variation like what has been reported for Standard Australian English.

Index Terms: CAAE, vowel, formant, acoustics, allophone

1. Introduction

Across Australia today, the majority of Aboriginal and Torres Strait Islander people speak English on a daily basis [1]. The English varieties spoken by Aboriginal people are often referred to as Australian Aboriginal English (AAE), and many of these varieties have been reported to diverge from Standard or Mainstream Australian English (SAE/MAE) in the phonetic and phonological properties of consonants and vowels. Some of these may reflect systematic differences between the varieties, and some of the differences may relate to whether an individual speaks English as a first or primary language (L1) or as a second or additional language (L2) [2]–[5]. For instance, some varieties of L2 AAE have a reduced vowel system with five vowels /ɪ ɛ ɐ ɔ ʊ/, or even three vowels /ɪ ɐ ʊ/, depending on the speaker's L1 vowel inventory, while SAE has thirteen monophthongs, including the vowels in FLEECE, KIT, SQUARE, DRESS, TRAP, NURSE, STRUT, BATH, LOT, THOUGHT, FOOT, GOOSE, and SCHWA (IPA: /iː ɪ eː e æ ɜː ɐ ɐː ɔ oː ʊ ʉː ə/) plus diphthongs [6]. In the present paper, we limit the scope to only monophthongs.

Central Australian Aboriginal English (CAAE) is a variety of English spoken by Aboriginal Australians living in Central Australia (Alice Springs/Mparntwe, Northern Territory) [7], [8]. Alice Springs is a highly multilingual community; early impressionistic reports have suggested that the phonetics of CAAE may have influence from local Indigenous Australian languages due to language contact. Among the many traditional Aboriginal languages spoken in and around Alice Springs, are Eastern and Central Arrernte, Western Arrarnta, Alyawarr, Anmatyerr, Kaytetye, Pintupi-Luritja, Pitjantjatjara, Yankunytjatjara, and Warlpiri.

The only previous study of the CAAE vowel system [9]—undertaken almost 20 years ago—examined the production of four multilingual speakers of English as a second language (L2); the speakers' first languages [L1s] were Eastern Arrernte, Warlpiri, or Western Desert language. The results from the study, which relied on vowels elicited in a /hVd/ context, suggested that the L2 CAAE speakers' vowel space is *phonologically* similar to SAE, as the phonemic vowels were acoustically differentiated from their neighbouring categories. No statistical tests were, however, carried out to check the significance of acoustic differences between categories.

In addition to indicating that CAAE shares the phonological system of SAE, the analysis in [9] also indicated that CAAE vowels differed *phonetically* from that of SAE speakers from South Australia. For example, the study showed that TRAP vowel /æ/ in CAAE was closer/higher in relation to the STRUT vowel /ɐ/, and the THOUGHT vowel /oː/ was lower in relation to the FOOT vowel /ʊ/ (whereas the two vowels had similar heights in SAE). Additionally, The GOOSE and NURSE vowels (/ʉː/ and /ɜː/) in CAAE seemed to be further back than SAE.

Further, since the experimental protocol in [9] relied on a /hVd/ word-reading task, it did not allow for analysis of potential allophonic variations in the realisation of the monophthongs in different phonological contexts as has been demonstrated for SAE [10]–[13]. As a consequence, it remains unclear whether this variety of AAE exhibits contextual allophony of the kind observed in SAE. At the same time, it remains unclear whether the current CAAE speakers, who use English as their primary language (L1), demonstrate similar acoustic vowel properties to the previous report based on L2 speakers.

Motivated by the scarcity and limitations in previous research, the present study provides an acoustic phonetic description of vowels in CAAE using naturalistic conversational speech data, including vowels spoken in a range of phonological contexts that have been reported to give rise to systematic allophonic variations in SAE [10]–[13]. Notably, /æ/ potentially has at least two allophones in SAE, namely TRAP and BAN (pre-nasal, tense); the /ɜː/ vowel has two allophones, NURSE and GIRL (pre-lateral). Finally, the /ʉː/ vowel has three allophones, the canonical GOOSE vowel, the TOO vowel (post-coronal) and the POOL vowel (pre-lateral).

The research questions of this paper are as follows:

(1) How does each CAAE phonemic vowel differ acoustically from other vowels?

(2) How do CAAE vowels occupy the acoustic vowel space, and whether CAAE vowels show significant differences in neighbouring pairs?

(3) Do CAAE vowels show contextual allophonic variations?

We approach the first question using a classification analysis, and for the second and third questions, we use acoustic phonetic analyses.

2. Methods

2.1. Participants

We recorded six female speakers of CAAE from Alice Springs, Northern Territory of Australia. The speakers ranged in age from 26-38 years (*M* = 32.7). All had learned English from childhood and reported speaking CAAE as their primary language. One woman additionally reported speaking the traditional Indigenous Australian languages of Eastern Arrernte and Alyawarr, and one also speaks the traditional Indigenous Australian languages of Anmetyerr and Pitjantjatjara regularly.

All participants were recruited by a senior English-Arrernte bilingual researcher involved in the *Little Kids Learning Languages* project (https://little-kids-learning-languages.net/), who visited community centres and residential areas in Alice Springs with another researcher from the project. All participants received a $AUD50 supermarket voucher for their participation.

The participants were recorded either in their homes (e.g., on the veranda) or at their local community centre, participating in conversational interviews of approximately 30 minutes duration, focused on everyday activities related to children and child-rearing and led by an Arrernte researcher fluent in Arrernte and CAAE. All interviews were recorded using a Sennheiser EW112 PG4-GB Portable Wireless Lapel Microphone System.

2.2. Data preparation

All recordings were transcribed and annotated in English orthography, and 1836 vowel tokens were manually labelled in the phonetics software Praat [14], using Standard Australian English (SAE) orthography [12]. We acknowledge that all orthographic transcriptions impose phonological distinctions on a dataset, but none-the-less take this approach under the assumption that SAE orthography captures the phonological distinctions of CAAE reflects a null hypothesis of sorts. Importantly, it allows us to make statistical comparisons between the formant and duration associated with each purported phonological category to reveal whether each SAE vowel contrast is also contrastive in CAAE.

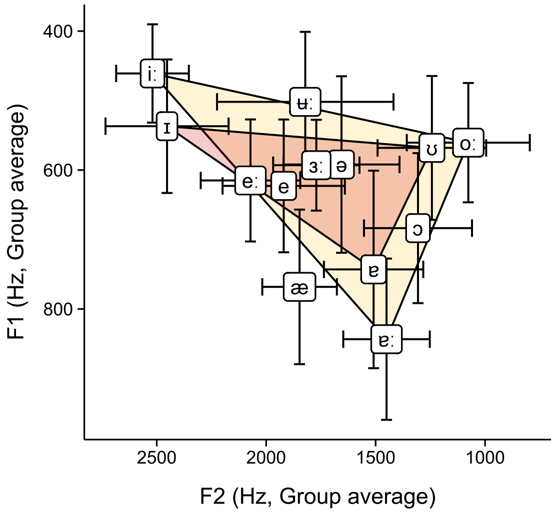


Figure 1. *Mid-point CAAE vowel qualities*

After annotation, we extracted spectral and durational measurements from each vowel, including the first four formants (F1-F4) and vowel duration using a customised *Praat* script. Formant values were estimated at the 10%, 30%, 50%, 70%, and 90% time points in the interval, to capture dynamic change over time. Erroneous data (e.g., zeroes) were excluded. For the vowel classification analysis, all formant data was included, but for the analysis of the acoustic vowel space, we included only the midpoint (50%) values which we assume represents the spectral quality of each vowel's steady state, defined by the first two formants, F1 and F2. Figure 1 presents the midpoint formant values for all CAAE phonemic vowels at the group level, where triangles indicate peripheral tense vowels (yellow) and lax vowels (red). Error bars indicate Standard Deviations. The mean duration values (in ms) are reported in Table 1.

Table 1. Mean durations (Dur.) in ms in CAAE vowels.

Vowel	Dur.	Vowel	Dur.	Vowel	Dur.
iː	106	eː	143	oː	136
ɪ	71	e	78	ʊ	67
ɐː	190	æ	108	ʉː	134
ɐ	80	ɔ	85	ɜː	153

2.3. Data analysis

We first conducted a phonemic vowel analysis (Section 3.1) to determine whether each pair of adjacent vowels also differ in CAAE, using a supervised learning algorithm (Random Forest, RM). Similar methods have been used in the other research of a similar kind, e.g., Linear Discriminant Analysis (LDA) as in [15], but RM has the additional advantage of allowing both numerical predictors (e.g., formants, duration) and categorical variables (e.g., individual speakers). Using the statistical software JASP, we fitted ten RM models based on the dataset, and then asked the models to perform predictions (70-30 split in training and testing, five-fold cross-validation).

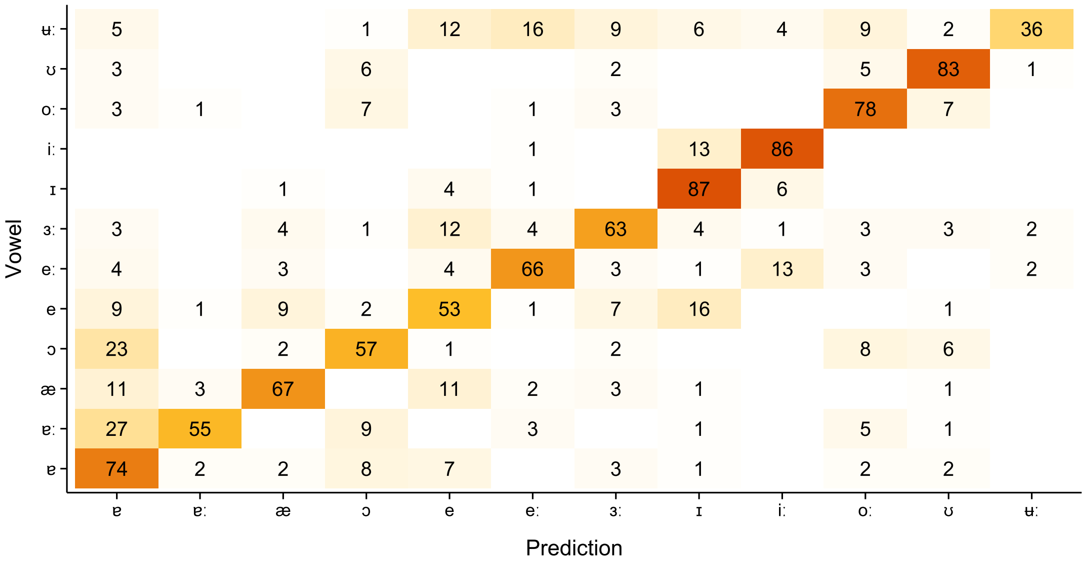
The classification results provided by the learning algorithm were used to generate an averaged classification matrix (Figure 2). We also analysed and visualised the CAAE vowels spoken by the speakers on a F1-F2 plane (Figure 3) in Section 3.2. Following conventional methods in acoustic phonetics, we then compared the acoustic parameters of the CAAE vowels focusing on five critical contrasts, including FLEECE-KIT (/iː/-/ɪ/), BATH-STRUT (/ɐː/-/ɐ/), SQUARE-DRESS (/eː/-/e/), THOUGHT-FOOT (/oː/-/ʊ/), and TRAP-DRESS (/æ/-/e/). This approach of course assumes that all such contrasts are maintained in CAAE.

Finally, we analysed the acoustic differences between the default allophones of CAAE vowels with their contextual allophones, e.g., TRAP-BAN, NURSE-GIRL, and GOOSE-TOO-POOL. We used linear mixed-effects modelling (LMM) for all statistical inferences (random intercepts for participants and all observed lexical items).

3. Results

3.1. Classification matrix

The RM classification models were trained with all acoustic measurements (F1-F4 at 10%, 30%, 50%, 70%, and 90% time points; duration) and speaker identity as predictors, and their average performance is summarised in Figure 2, where CAAE



Vowel	ɐ	ɐː	æ	ɔ	e	eː	ɜː	ɪ	iː	oː	ʊ	ʉː
ʉː	5			1	12	16	9	6	4	9	2	36
ʊ	3			6			2			5	83	1
oː	3	1		7		1	3			78	7	
iː						1		13	86			
ɪ			1		4	1		87	6			
ɜː	3		4	1	12	4	63	4	1	3	3	2
eː	4		3		4	66	3	1	13	3		2
e	9	1	9	2	53	1	7	16			1	
ɔ	23		2	57	1		2			8	6	
æ	11	3	67		11	2	3	1			1	
ɐː	27	55		9		3		1		5	1	
ɐ	74	2	2	8	7		3	1		2	2	

Figure 2. *Classification matrix of CAAE vowels (Random Forest models)*

vowel targets are shown on the y-axis, and the predicted categories are shown on the x-axis. Only phonemic vowels were included in this analysis, and the contextual allophones (BAN, GIRL, TOO, and POOL) were excluded.

The cells on the diagonal line display percentages of time when correct classification was performed, i.e., the predicted category matched the target category, where the chance-level is 1/12 (8.3%). Among the 12 phonemic vowels (excluding the unstressed vowel /ə/), high classification accuracies were observed for the peripheral vowels in KIT (87%), FLEECE (86%), FOOT (83%), THOUGHT (78%), and STRUT (74%), indicating that these five vowels have low levels of acoustic overlap with neighbouring CAAE vowels. Relatively lower accuracies were seen for TRAP (67%), SQUARE (66%), NURSE (63%), LOT (57%), BATH (55%), and DRESS (53%), indicating that these vowels potentially show a certain level of acoustic overlap with neighbouring vowels (see also Figure 1). Finally, we observed the lowest accuracy measure for GOOSE (36%), as the GOOSE targets were often misclassified into a wide range of vowel categories. However, other categories are seldom misclassified as GOOSE (< 5%), indicating that the vowel remains distinct from other categories. These results also show that misclassifications are not necessarily symmetrical.

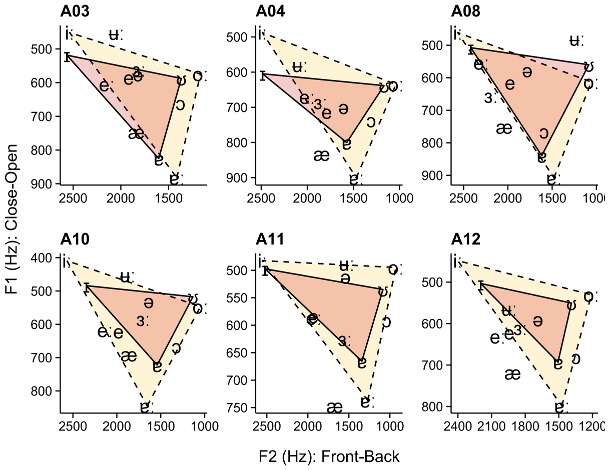


Figure 3. *Individual CAAE vowel plots*

3.2. CAAE vowel space

Figure 3 presents the acoustic vowel space of each CAAE speaker. Again, we visualised the triangles defined by the centroids of peripheral vowels, both tense and lax, (FLEECE-BATH-THOUGH, and KIT-STRUT-FOOT). All speakers produced all phonemic vowels in CAAE, but individual differences also existed in terms of how each vowel differed spectrally from those neighbouring categories. Here, we are primarily interested in the acoustic differences in critical (acoustically and articulatorily adjacent) pairs. The *p*-values reported below are the results of *F*-tests based on LMMs (mixed effects: *speaker* and *lexical item*).

For the FLEECE-KIT contrast, our analysis indicated that FLEECE vowels had longer durations ($p < .0001$), lower F1s ($p < .0001$), and higher F2s ($p = .0001$) than KIT vowels. The difference in F2 was more prominent in speakers A08, A10, and A12, as compared to other participants. With respect to the BATH-STRUT contrast, our analysis showed that BATH vowels had longer durations ($p < .0001$) and lower F1s ($p = .0164$) than STRUT vowels, but the BATH and STRUT vowels had similar F2s ($p = .2348$). The F1 difference was present in all six speakers. For the final tense-lax pair—SQUARE-DRESS—the analyses showed that SQUARE vowels had longer durations ($p < .0001$) and higher F2s ($p = .0172$) than DRESS vowels, but the difference in F1 was not significant ($p = .8175$). The difference in F2 was prominent in A03, A04, A08, and to a lesser extent A10 and A12, but potentially not A11. Finally, we compared between TRAP and DRESS, and the analysis showed that TRAP vowels had longer durations ($p = .0040$), and higher F1s ($p < .0001$) than DRESS vowels, but the two vowels had similar F2 values ($p = .6337$).

Our acoustic analysis also indicated that the THOUGHT and FOOT vowels were high back vowels in CAAE, as in SAE, and that the two vowels had similar F1 values ($p = .1062$) as well as F2 values ($p = .3220$), but THOUGHT vowels were longer than FOOT vowels ($p = .0002$), as expected. Lastly, the results indicated that that GOOSE vowels were produced with substantial fronting, as a central vowel, by all CAAE speakers, except for A08, who produced the vowel as a high back vowel (even higher than THOUGHT). Additionally, TRAP vowels ere produced with a similar vowel height to STRUT vowels, but substantial individual differences were also observed, e.g., sometimes the TRAP vowel was higher than the STRUT vowel (as in A03, A08), and sometimes it was lower than the STRUT vowel (as in A04, A11).

3.3. Contextual allophones

In what follows, we compare the acoustic quality of the contextual allophones with their canonical, or 'elsewhere' allophones. The relative position of these allophones is displayed in Figure 4, for each CAAE speaker. Note that some speakers did not produce all allophones, as a consequence of the naturalistic and conversational data collection.

We first consider the TRAP vowel and its pre-nasal allophone, the BAN vowel. Here our tests showed that BAN vowels had higher F2 values than TRAP vowels (*p* = .0011), and BAN vowels were also potentially longer than TRAP vowels (*p* = .0807), but these two had similar F1 values overall (*p* = .2900). In summary, BAN vowels tended to be longer, and more fronted than the default allophone, the TRAP vowel.

With respect to the NURSE vowel and its pre-lateral allophone, the GIRL vowel, the tests showed that the two vowels had similar durations (*p* = .8652), similar F1 values (*p* = .5096), but that the GIRL vowel potentially had lower F2 values (*p* = .0836). The F2 difference was more prominent in A04, A11, and A12, but not in other speakers. Finally, we consider the GOOSE vowel and its post-coronal allophone, TOO, as well as its pre-lateral allophone, POOL. Here, the tests showed that the three allophones did not differ statistically in duration (*p* = .3421), or F1 (*p* = .2799), though a significant difference was found in F2 values (*p* < .0001). A series of *post hoc* tests revealed that the POOL vowel had lower F2 values than both the GOOSE and TOO vowels (*p* < .0001 for two tests), but the difference between the GOOSE and TOO vowels was not significant (*p* = .2122). Therefore, our dataset indicates that the /ʉː/ vowel is more influenced by the lateral context than the coronal segments preceding it in CAAE than in SAE/MAE.

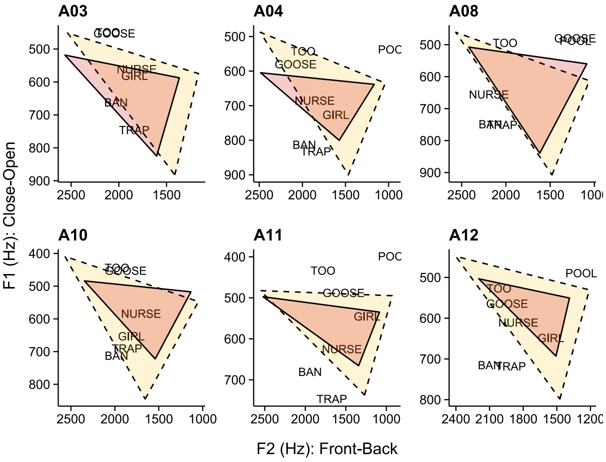


Figure 4. *Individual CAAE allophones*

4. General Discussion

The present study focused on characterising the structure of the monophthong inventory of CAAE. The results clearly indicate that the phonological structure of CAAE monophthongs is like the system in SAE, but also reveal phonetic differences in the realisation of CAAE vowels relative to SAE.

Our RM classification analysis shows that most vowels achieve an accuracy of higher than 50% in a twelve-alternative classification task with a chance-level of only 8.3%. The highest accuracy measures (>70%) were observed for peripheral vowels, presumably because these vowel categories do not share acoustic spaces with many other categories; conversely, accuracy measures were lower (>50%) in most non-peripheral vowels, potentially because they share acoustic spaces with multiple neighbouring categories. At the same time, lax vowels could be perceived/categorised as undershoots of tense vowels. Finally, the GOOSE vowel was often misclassified into other vowels, but other vowels were seldom misclassified as the GOOSE vowel, displaying an asymmetrical pattern. This phenomenon is potentially due to a large within-category variability of the GOOSE vowel over other neighbouring vowels, and more research is required to further investigate the role of relative dispersion level in vowel misclassification. In general, our RM classification suggests that all vowels were phonemically different in L1 CAAE, consistent with the previous report of L2 speakers based on a /hVd/ word-reading task [9].

Further acoustic comparisons between neighbouring pairs (Figure 2) also confirmed the finding. For example, the tense-lax pairs (e.g., BATH-STRUT, FLEECE-KIT, SQAURE-DRESS) all showed a clear difference in vowel duration, but spectral qualities also played an important role in these pairs, e.g., the BATH vowel was often lower than the STRUT vowel, and the FLEECE vowel was often more front and closer/higher than the KIT vowel. Additionally, the SQAURE vowel was sometimes more front than the DRESS vowel. These spectral features potentially differ from SAE documented in the literature, where the spectral differences between these three pairs were minimal [16]. In general, in the tense-lax pairs of CAAE, the tense vowels tend to be more peripheral than the lax vowels. See also that the red triangles were often enclosed in the yellow triangles (Figure 2). Additionally, we confirmed that a contrast between the TRAP and DRESS vowels was maintained by duration and F1, and a contrast between the THOUGHT and FOOT vowels was maintained by duration.

Finally, this study presents a first analysis of vowel allophony in CAAE, and tested whether the common allophones in MAE/SAE [12] also existed in this Aboriginal variety. Indeed, we observed that /æ/ was fronted and potentially lengthened in a pre-nasal context (i.e., the BAN vowel) as compared to its default allophone TRAP. For the /ɜː/ vowel our analyses showed that a pre-lateral context potentially led to lower F2 values, but the difference we found was nuanced and it did not reach the significance level. Finally, the /ʉː/ vowel is by default a central vowel in CAAE, as in SAE, but it can become a back vowel in a pre-lateral context (i.e., POOL), while the influence of a preceding coronal consonant was unclear, as we did not observe significant differences between TOO and GOOSE. These effects could be further investigated using a more controlled experiment-based design, potentially also with SAE control groups.

In conclusion, the study contributes to the small set of instrumental investigations into the phonology and phonetics of Australian Aboriginal varieties of English, and thus contributes to our understanding of the linguistic variations across communities in Australia. The study also highlights the benefits of using spontaneous conversational speech data as such data allows for tapping into the natural variation of any given language variety in ways that word-list or nonse-word elicitation, such as the /hVd/-list implemented in [9] might not.

5. Acknowledgements

We thank the participants in the study, Tangentyere Council Town Camps Community Centres, Charles Darwin University, Central Australian Aboriginal Congress and Batchelor Institute.

6. References

[1] ABS, “Language statistics for Aboriginal and Torres Strait Islander peoples,” 2021. [Online]. Available: https://www.abs.gov.au/statistics/people/aboriginal-and-torres-strait-islander-peoples/language-statistics-aboriginal-and-torres-strait-islander-peoples/latest-release

[2] M. Sharpe, “Alice Springs Aboriginal Children’s English,” in *Australian linguistic studies*, S. A. Wurm, Ed. Canberra, Australia, Australia: Pacific Linguistics, The Australian National University, 1979, pp. 733–748. doi: 10.15144/PL-C54.733.

[3] I. G. Malcolm, “Australian creoles and Aboriginal English: phonetics and phonology,” in *Varieties of English 3: The Pacific and Australasia*, K. Burridge and K. Bernd, Eds. Berlin & New York: Mouton de Gruyter, 2008, pp. 124–141.

[4] A. Butcher, “Linguistic aspects of Australian Aboriginal English,” *Clin. Linguist. Phonetics*, vol. 22, no. 8, pp. 625–642, 2008, doi: 10.1080/02699200802223535.

[5] J. Harkins, *Bridging two worlds: Aboriginal English and crosscultural understanding*. Brisbane, Australia: University of Queensland Press, 1994.

[6] J. C. Wells, *Accents of English: Volume 1*. Cambridge: Cambridge University Press, 1982.

[7] H. Koch, “Central Australian Aboriginal English: In Comparison with the morphosyntactic categories of Kaytetye,” *Aisan Englishes*, vol. 3, no. 2, pp. 32–58, 2000, doi: 10.1080/13488678.2000.10801054.

[8] H. Koch, “The influence of Arandic languages on Central Australian Aboriginal English,” in *Creoles, their substrates, and language typology*, C. Lefebvre, Ed. John Benjamins Publishing Company, 2011, pp. 437–460.

[9] A. Butcher and V. Anderson, “The vowels of Australian Aboriginal English,” in *Proceedings of INTERSPEECH 2008*, 2008, pp. 347–350. doi: 10.21437/Interspeech.2008-145.

[10] J. Harrington, F. Cox, and Z. Evans, “An acoustic phonetic study of broad, general, and cultivated Australian English vowels,” *Aust. J. Linguist.*, vol. 17, no. 2, pp. 155–184, 1997, doi: 10.1080/07268609708599550.

[11] F. Cox and S. Palethorpe, “Australian English,” *J. Int. Phon. Assoc.*, vol. 37, no. 3, pp. 341–350, 2007, doi: 10.1017/S0025100307003192.

[12] B. Purser, J. Grama, and C. E. Travis, “Australian English vower time: Using sociolinguistic analysis to inform dialect coaching,” *Voice Speech Rev.*, vol. 14, no. 3, pp. 269–291, 2020, doi: 10.1080/23268263.2020.1750791.

[13] F. Cox, “Vowel change in Australian English,” *Phonetica*, vol. 56, no. 1–2, pp. 1–27, 1999, doi: 10.1159/000028438.

[14] P. Boersma and D. Weenink, “Praat, a system for doing phonetics by computer [Computer program],” *Version 6.3*, 2023, [Online]. Available: https://www.praat.org

[15] J. M. Hillenbrand, “Vowel classification based on fundamental frequency and formant frequencies,” *J. Speech, Lang. Hear. Res.*, vol. 36, no. 4, pp. 694–700, 1993, doi: 10.1044/jshr.3604.694.

[16] F. Cox and J. Fletcher, *Australian English: Pronunciation and transcription*. Melbourne, Australia: Cambridge University Press, 2017.

A Database of Multilingual Child Speech with Recordings from a Longitudinal Project for Multilingual Education

Paola Escudero[1], *Gloria Pino Escobar*[1], *Milena Hernández Gallego*[1], *Chloé Diskin-Holdaway*[2] and *John Hajek*[2]

[1] MARCS Institute for Brain, Behaviour and Development, Western Sydney University, New South Wales, Australia, [2]School of Languages and Linguistics, The University of Melbourne, Victoria, Australia

paola.escudero@westernsydney.edu.au, g.pinoescobar@westernsydney.edu.au, m.hernandezgallego@westernsydney.edu.au, chloe.diskinholdaway@unimelb.edu.au, j.hajek@unimelb.edu.au

Abstract

We introduce a project focused on developing a multilingual database of Australian children's English and Heritage Language (HL) speech, starting with Spanish and adding other HLs progressively. Leveraging the technology and expertise from the AusKidTalk corpus [1], our database currently comprises approximately 610 hours of speech from children aged 3-7 years who speak English only or English and Spanish. Data were collected through online testing sessions featuring eight psycholinguistic tasks designed to elicit both single-word, sentence, and short story productions. This paper outlines the key features, design, data collection and analysis method, as well as the repository storage for data management. The aim is to facilitate linguistic research on language development in monolingual and multilingual Australian children. In this paper, we showcase the database, discuss the analyses conducted so far, and outline projects that have already used the database as well as future related projects. Additionally, we will detail our current data management plan and share our vision for collaborating with the broader research community.

Index Terms: corpus linguistics, speech corpus, children's speech, Spanish-English bilinguals, child bilinguals.

1. Introduction

The understanding, modeling, and conceptualizing of language acquisition and development relies heavily on language datasets and corpora [2, 3]. Despite interest in the analysis of natural language use increasing in the last decade (see studies listed in [4]), there is still a shortage of adult and child speech corpora worldwide, including in Australia.

Child speech corpora are especially difficult to find, with currently fewer than 20 worldwide (exceptions include CHILDES and AusKidTalk [1, 5]). This may be attributed to the fact that child speech, particularly that of young children, is relatively difficult to collect and analyse [1]. Their scarcity notwithstanding, child corpora offer numerous advantages for the study of child bilingualism (see [2] for a more detailed description). For instance, the CHILDES corpus has facilitated more than 1,300 studies on language disorders, second language acquisition, literacy, among other disciplines [5].

Within the existing speech corpora, spoken ones remain a minority compared to written ones [6], with many more monolingual compared to multilingual corpora. Exceptions include the CHILDES Russian-German ZAS-MAIN Corpus [7] and the Leibniz-ZAS corpus of MAIN [8]. Importantly, the unpredictability of spontaneous child speech has led most child corpora to employ constrained protocols with limited tasks, and few are fully transcribed and annotated [9].

The current database aims at including fully annotated and transcribed child speech data resulting from recording sessions. It features eight different tasks, which could help the research community to answer a variety of linguistic and psycholinguistic empirical questions. The recordings of these psycholinguistic tasks are part of the data collected within a large longitudinal research program which aims at fostering multilingualism through play-based language immersion sessions in preschools [10, 11]. This longitudinal research program also includes recordings of language exposure sessions, electronic forms used to evaluate children's performance in those sessions, with results reported in in [10]. Here we report on the speech database from the psycholinguistic tasks in English and Spanish due to space limitations.

Our child speech database is intended to make a significant contribution to the study of child bilingualism as the first multilingual corpus of child speech in English and heritage languages (HLs) in Australia. It aims at filling in the gap in child speech corpora mentioned above with detailed documentation of how children's languages other than English (LOTEs) develop, which are learned alongside this societal language. This database is an invaluable tool to conduct research on children's linguistic, cognitive, and socio-emotional development, including collaboration among researchers with multidisciplinary expertise. The main characteristics of the data collection design and analysis which have been used in the first five articles written using these database [10-14] are as follows:

1. Online data collection for Australian preschoolers in English and Heritage Languages,
2. Data de-identification, transcription of recordings, curation and connection to linguistic background and demographic information,
3. Ongoing transcription of all collected data using available computational tools that enable speech-to-text for orthographic and phonemic segmentation of spoken utterances.
4. Availability of data to other researchers by writing to the first author and data custodian.

Below we explain the specific features of the corpus data collection, including participants and testing protocol, which we hope to continue as more data is collected. The goal is to streamline both collection and transcription as data entries are curated using the same effective and efficient procedure.

2. Method

The part of the database reported here includes child speech that allows the evaluation of multiple linguistic features in a single, child-friendly recording session. Different utterance types, e.g., single words, sentences, and narratives, were elicited to facilitate research into monolingual and bilingual language production and development. Crucially, demographic metadata that were collected for each participant have been linked to the recordings for essential understanding of each child's unique language experience.

The eight tasks were conducted online in a single Zoom session which lasted for approximately 45 minutes (session duration range: 45-60 mins), reducing the likelihood of participant dropout. The recording procedure enabled easier participation than in most previous infant and child experiments that take place in laboratories, as families are unlikely to readily travel long distances to university-based laboratories. Importantly, the suitability and reliability of using audio recordings made with Zoom for linguistic studies has been confirmed by [15-17]. Tasks were presented to parents and children as games and activities with the aim of eliciting different types of speech samples while maintaining children's engagement during the session (11-15).

Parents were asked to complete a Qualtrics online survey which included important demographic and sociolinguistic information, which is indispensable for linguistic analysis because failure to connect data to metadata would lead to “nothing but disconnected words of unknowable provenance or authenticity” [18]. We thus ensured that all data entries were correctly indexed with their corresponding child speech data (point 5), and extra data checks were conducted for verification.

2.1 Participants

As shown in Table 1, the current data set includes sessions from 64 children who resided in Australia, with no diagnosis of language or developmental disorder (M age = 4.60 years, range 3-7). They were recruited from a database of parents who had volunteered to participate in research at an Australian university laboratory (n = 27) and from a bilingual preschool located in Sydney, Australia, where parents had volunteered to participate in the larger longitudinal program mentioned above [10] for HL maintenance and additional language (AL) learning (n = 37). Of the participants, 18 were English monolinguals, 39 were Spanish-English bilinguals, and 7 were bilinguals of English and another language. The Spanish-English bilingual group was further divided into HL simultaneous bilinguals (n = 19), who had acquired Spanish at home, and children who had acquired Spanish as an AL in their childcare setting (n = 20).

Following the recruitment procedure from previous studies [11,15], parents received an initial email with a link to the study's information sheet, consent form, and demographic survey, all hosted on Qualtrics (https://www.qualtrics.com). Participation was voluntary, with parents providing written consent and children giving oral assent prior to their participation. Parents completed the demographic questionnaire after consenting online. The survey included questions about the child's language background, such as weekly language exposure and use of English, Spanish, and any additional languages, along with their daily routines. Caregivers also reported their own language background, proficiency, and language use with their child. Parents then received a follow-up email with instructions to schedule their child's online session via Calendly (https://calendly.com/). The study was approved by the Western Sydney University Human Research Ethics Committee (approval number: H11022), with the first author serving as the data custodian. All data was anonymized to protect children's identities.

122 files with a duration of 45-60 minutes each have been organized and prepared for transcription to date. There are currently 67 recordings in English and 55 recordings in Spanish. The demographic survey shows that all the multilingual children were acquiring two languages, and some three, as shown in a recent study reporting on one of the psycholinguistic tasks with the same participants [11].

Table 1: *Parental report in the demographic survey for the participants included in the corpus (n=64)*

	English mono-lingual	Other bilingual	Spanish-English HL	Spanish as an AL
N	18	7	19	20
Mean age (range)	4.4 (3-5)	4.4 (3-5)	4.7 (3-7)	4.8 (4-6)
Mean English exposure % (range)	97.40 (70.7-100)	45.13 (5.2 -65.0)	42.9 (14-80)	75.8 (45-100)
Mean 1st LOTE exposure % (range)	2.4 (0-29.0)	47.43 (14.0-95)	52.9 (20-86)	21.8 (0-52)
Median Principal Carer Relation	Mother (n=14)	Mother (n=6)	Mother (n=18)	Mother (n=19)
Principal Carer Education	University degree 85%	University degree 85%	University degree 85%	University degree 85%

The research project was designed as a pre- and post-intervention study, and the children's proficiency was measured both in Spanish and English, over three timepoints: two prior (2021) and one posterior (2022), to the delivery of Spanish language immersion sessions reported in [10], as the children transitioned from preschool to primary school. Given that some participants did not complete all pre and post sessions, some researchers may consider the data pseudo-longitudinal (cf. Corpus of Learner German (CLEG13) [2]) and cross-sectional.

The current dataset includes 92 recordings resulting from the first timepoint of data collection (45 in Spanish and 47 in English). 22 participants were tested at both the first and the second timepoint (15 in Spanish and 7 in English). Many participants have continued their participation in the longitudinal project after graduating from preschool, with data from these subsequent sessions awaiting organisation and connection to their correspondent demographic information. The dataset mainly includes Spanish and English sessions from the first session, as most participants were only tested in Spanish because their English proficiency was comparable to that of monolingual peers [11]. Additionally, the aim of the longitudinal project was to investigate the development of HL proficiency, due to the demonstrated difficulty in maintaining HL proficiency in Australia [19].

2.2 Online data collection protocol

Recordings were conducted online following the online testing procedures described in [11, 15]. The sessions took place via Zoom and a link was sent to parents prior to the session for use with a home device (preferably a laptop or desktop), with an experimenter guiding the children through eight tasks assisted by one of the child's parents for session setup [15]. The tasks included an initial exposure to an audiovisual eBook, followed by a retelling and comprehension task [11-14], a digit span task [20], a nonword repetition task [21], two verbal fluency tasks [22], a receptive vocabulary task [23], and an episodic memory task [24], in that order. Each session was recorded for response data reliability [15, 16] and for speech database building.

The audiovisual eBook was a 12-page electronic storybook featuring colourful 2D line drawings, which was narrated by a female native speaker of Australian English. The story depicted two children sharing fruit and toys at school. After exposure to the eBook, children completed a retelling task where they recounted the story in their own words [23, 24], followed by a comprehension task that involved answering questions with visual aids [11, 14, 25, 26]. The retelling task was always presented first to avoid bias from comprehension questions, as studies have shown that comprehension does not influence retelling accuracy [27, 28]. The digit span task measured short-term verbal memory with children asked to recall sequences of numbers that increased in length [20]. The nonword repetition task measured phonological memory as the ability to store and reproduce unfamiliar sound sequences in the target language [21]. The verbal fluency tasks assessed lexical access and cognition by asking participants to retrieve words from either a specific category or beginning with a particular sound, within a 1-minute time limit [22]. In the receptive vocabulary task, children listened to a word and identified its corresponding image, demonstrating their understanding of the word's meaning [23]. Finally, in the episodic memory task, children recalled a sequence of actions previously presented, such as the steps involved in celebrating a birthday [24]. These eight tasks provided speech data in the target language.

The present corpus did not involve additional specialised recording apparatus as the data was obtained via online Zoom recordings following previous studies [11, 15]. Using the Zoom subscription that most universities have, our online procedure represents a time and cost-efficient option for large-scale longitudinal projects such as the one in [11].

3. Speech data transcription, analysis, organisation, and storage

So far, we have partially transcribed and annotated the current dataset, using the following procedure: the audio recordings sampled at 16kHz or above were first extracted as WAV files from Zoom sessions run on desktop/laptop computers. These files underwent orthographical transcription at the utterance level using OpenAI's Whisper [29] an automatic speech recognition (ASR) system that is trained on approximately 680,000 hours of diverse, multilingual data sourced from the internet. Transcription can be done much more quickly and with fewer human resources when ASR tools such as Whisper are used.

The text transcriptions, linguistic variable labelling, and time information were then converted into Praat TextGrid files using MATLAB scripts. Research assistants manually verified the correctness of these transcriptions. Subsequently, the audio and transcription data underwent forced alignment using WebMAUS [30] with the Australian English model, to implement annotation at both the word and phoneme levels. The research team manually verified these boundaries and annotations, adhering primarily to the automated suggestions unless errors were evident. Aside from time efficiency. compared with manual transcription, Whisper ASR followed by manual verification substantially decreases the human resources required for transcription. OpenAI technology to transcribe the recorded data thus facilitated automated pre-processing and segmentation, and reduced the time required for manual phoneme-level annotation. Units of speech such as words, utterances, and errors at the word level were further annotated for phonetic, phonological, semantic, and morphological analysis [31].

For phonetic analysis, research assistants also labelled segments and syllables depending on the specific topic, e.g., stressed and unstressed syllables [13] or types of voiceless plosives [14]. All scripts and tables generated from this segmentation procedure are part of the documentation and metadata of the database with fully de-identified data that can be made available to researchers via the project's principal investigator and data custodian (first author of this paper).

All files were then named including date of collection, length of the file, participant type, ID number, and age in the file name, following the standard file naming system for the Talkbank and CHILDES (Child Language Data Exchange System) projects, called the Codes for the Human Analysis of Transcripts (CHAT) [32] and indexed in a Microsoft Excel spreadsheet. All data have been stored along with their corresponding metadata in a WSU online data repository, with full access for the current research team, and with future access available through the principal investigator and data custodian. Additionally, this data repository ensures that sensitive data are protected by security safeguards against loss, unauthorised access, misuse, or disclosure. The five articles and any new articles using parts of the dataset will be connected to the data repository.

Our transcription procedure has been followed for the already reported studies using this dataset [11-14]. We will finish transcription and annotation of the recordings for the 64 participants and 122 files following the above-mentioned procedure and included in the current corpus. Further data collected within the longitudinal study will follow the same procedure. As was mentioned in a previous presentation of the database, it includes data collected in language exposure sessions for HL and AL children that have been delivered in French, Mandarin, Spanish, and Vietnamese [33], which are part of the longitudinal research program [10]. Additionally, ongoing research will report on psycholinguistic data including datasheets with participant scores from the Spanish retelling and comprehension tasks, the memory, verbal fluency, and sequential memory tasks, along with the already included dataset with the results of the English retelling and comprehension task reported in [11].

4. Discussion and Conclusions

This project makes a significant contribution to the study of child bilingualism and language acquisition by developing a multilingual and multimodal database of Australian children's English and Spanish speech. This database has already allowed for a broad range of research topics across multiple linguistic domains, such as phonetics, phonology, lexical acquisition, and narrative discourse, including cross-linguistic comparisons. The variety of tasks included, such as story retelling, auditory and phonological memory, and verbal fluency, among others, ensures that researchers can explore different aspects of language development, from sentence structure to vocabulary acquisition. By incorporating demographic metadata, the database offers a robust contextual framework for examining individual differences in bilingual language acquisition. This comprehensive approach enables a deeper understanding of how factors such as age, language exposure, and social context shape linguistic development, with opportunities for other researchers to harness this important resource.

The use of online experimental sessions and modern transcription technologies, such as OpenAI, highlights the efficiency and innovation of the project. By providing de-identified, fully transcribed and annotated data, the database saves researchers time and reduces transcription biases, allowing for more accurate analyses from diverse theoretical perspectives. The project's commitment to transparency and accessibility through the data custodian sets a strong precedent for future linguistic research, promoting collaboration and open science practices, while at the same time protecting participants' anonymity.

In addition to its focus on child bilingualism, this corpus contributes to the broader research agenda on LOTEs in Australia and beyond. It provides valuable insights for educational strategies and language learning programs, and it holds potential for informing clinical interventions for children with language disorders. The detailed speech data can also help improve automatic speech recognition (ASR) systems, particularly in recognizing speech patterns in spontaneous conversational settings in young multilingual populations, thus addressing a key technological challenge [1, 34].

5. Acknowledgments

This study was supported by three grants from the Australian Research Council awarded to the first author (CE140100041, FT160100514) and to the first and two last authors (LP210300631). Special thanks go to Dr. Weicong Li for his assistance with automatic transcription and initial and ongoing database organization, and to our dedicated team of research assistants for their invaluable help with data collection and the manual refinement of automatic transcription. We extend our sincere thanks to the children who participated in this study and to their parents for providing consent.

6. References

[1] B. Ahmed et al., "AusKidTalk: an auditory-visual corpus of 3-to 12-year-old Australian children's speech", Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, 3680-3684, 2021,

[2] S. Granger, G. Gilquin, and F. Meunier, The Cambridge Handbook of Learner Corpus Research. Cambridge University Press, 2015.

[3] B. MacWhinney, "Understanding spoken language through TalkBank," Behavior Research Methods, vol. 51, no. 4, pp. 1919–1927, Dec. 2018.

[4] M. Planelles Almeida, J.A. Duñabeitia and A. Doquin de Saint Preux (2022). "The VIDAS Data Set: A Spoken Corpus of Migrant and Refugee Spanish Learners". Frontiers in Psychology, vol. 12, Jan. 2022.

[5] B. MacWhinney, The Childes Project. Psychology Press, 2014

[6] J. Fernández and T.R. Davis, "Overview of Available Learner Corpora" in The Routledge Handbook of Second Language Acquisition and Corpora, N. Tracy-Ventura and M. Paquot,1st Ed, New York, U.S. Routledge, 2021, Country: Abbrev. of Publisher, year, pp. 147-159.

[7] N. Gagarina, "Narratives of Russian–German preschool and primary school bilinguals: Rasskaz and Erzaehlung," Applied Psycholinguistics, vol. 37, no. 1, pp. 91–122, Dec. 2015

[8] N. Topaj, Z. Rizaeva, A. Sternharzy N. Gagarina, "Leibniz-ZAS corpus of MAIN". Zenodo, abr. 28, 2021.

[9] N. F. Chen, R. Tong, D. Wee, P. Lee, B. Ma, and H. Li, "SingaKids-Mandarin: Speech Corpus of Singaporean Children Speaking Mandarin Chinese," In Proceedings of Interspeech 8 Sep. 2016.

[10] P. Escudero, G. Pino Escobar, C. Diskin-Holdaway, and J. Hajek. "Enhancing Heritage and Additional Language Learning in the Preschool Years: Longitudinal Implementation of the Little Multilingual Minds Program." OSF Preprints. September 16, 2024. https://doi.org/10.31219/osf.io/rvjc

[11] G. Pino Escobar and P. Escudero. "Vocabulary, Comprehension and Retelling in Multilingual Children: Age and Input Tell the Whole Story." OSF Preprints. September 16, 2024. https://doi.org/10.31219/osf.io/8e4nf

[12] M. Hernández Gallego, G. Pino Escobar, P. Escudero. "English lexical productivity and diversity in Spanish-English bilingual children in Australia." Proceedings of the 19th Australasian International Conference on Speech Science and Technology, SST 2024.

[13] Escudero, P., Li, W & Diskin-Holdaway, C. Have four-year-olds mastered vowel reduction in English? An acoustic analysis of bilingual and monolingual child storytelling. Proceedings of the 19th Australasian International Conference on Speech Science and Technology, SST 2024.

[14] Diskin-Holdaway, C., Li, W & Escudero, P. Bilingual preschoolers' phonetic variation keeps up with monolingual peers: The case of voiceless plosives in Australian English. Proceedings of the 19th Australasian International Conference on Speech Science and Technology, SST 2024.

[15] P. Escudero, G. Pino Escobar, C. G. Casey & K. Sommer, "Four-year-old's online versus face-to-face word learning via eBooks", Frontiers in Psychology, 12, 450, 2021.

[16] C. Zhang, K. Jepson, G. Lohfink, & A. Arvaniti, A, "Speech data collection at a distance: Comparing the reliability of acoustic cues across homemade recordings", The Journal of the Acoustical Society of America, 148(4_Supplement), 2717-2717, 2020.

[17] C. Ge, Y. Xiong, & P. Mok, "How Reliable Are Phonetic Data Collected Remotely? Comparison of Recording Devices and Environments on Acoustic Measurements", Proceedings of Interspeech, pp. 3984-3988, 2021.

[18] L. Burnard, "Metadata for corpus work", in M. Wynne, Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbow Books, pp. 30–46, 2005.

[19] P. Escudero, C. Jones Diaz, J. Hajek, G. Wigglesworth, and E. A. Smit, "Probability of heritage language use at a supportive early childhood setting in Australia," *Frontiers in Education,* vol. 5, p. 93, 2020, doi: 10.3389/feduc.2020.00093.

[20] S. E. Gathercole, "The assessment of phonological memory skills in preschool children," British Journal of Educational Psychology, vol. 65, no. 2, pp. 155-164, 1995.

◀◀ ◀ Table of Contents for this manuscript ▶ ▶▶

[21] S. E. Gathercole, “Nonword repetition and word learning: The nature of the relationship,” Applied Psycholinguistics, vol. 27, no. 4, pp. 513-543, 2006.

[22] M. Regard, E. Strauss, and P. Knapp, “Children’s production on verbal and non-verbal fluency tasks,” Perceptual and Motor Skills, vol. 55, no. 3, pp. 839-844, 1982.

[23] S. Weintraub, S. S. Dikmen, R. K. Heaton, D. S. Tulsky, P. D. Zelazo, P. J. Bauer, et al., "Cognition assessment using the NIH Toolbox," Neurology, vol. 80, no. 11 Supplement 3, pp. S54-S64, 2013.

[24] S. S. Dikmen, P. J. Bauer, S. Weintraub, D. Mungas, J. Slotkin, J. L. Beaumont, R. Gershon, N. R. Temkin, and R. K. Heaton, "Measuring episodic memory across the lifespan: NIH Toolbox Picture Sequence Memory Test," Journal of the International Neuropsychological Society, vol. 20, no. 6, pp. 611-619, 2014.

[25] J. Heilmann, J. F. Miller, A. Nockerts, and C. Dunaway, “Properties of the Narrative Scoring Scheme Using Narrative Retells in Young School-Age Children,” American Journal of Speech-Language Pathology, vol. 19, no. 2, pp. 154–166, May 2010, doi: https://doi.org/10.1044/1058-0360(2009/08-0024).

[26] N. Gagarina et al., “Assessment of Narrative Abilities in Bilingual Children,” Multilingual Matters eBooks, pp. 243–276, Dec. 2015.

[27] K. Kawar, E. Saiegh-Haddad, and S. Armon-Lotem, “Text complexity and variety factors in narrative retelling and narrative comprehension among Arabic-speaking preschool children,” First Language, p. 014272372211498, Feb. 2023.

[28] M. Silva and K. Cain, “The use of questions to scaffold narrative coherence and cohesion,” Journal of Research in Reading, vol. 42, no. 1, pp. 1–17, Oct. 2017.

[29] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning (pp. 28492-28518). PMLR, July, 2013.

[30] T. Kisler, U. Reichel, and F. Schiel, “Multilingual processing of speech via web services”, Computer Speech & Language, vol. 45, pp. 326–347, 2017.

[31] B. MacWhinney, “The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs”. Psychology Press, 2014.

[32] K. E. Squires, M. J. Lugo-Neris, E. D. Peña, L. M. Bedore, T. M. Bohman, and R. B. Gillam, “Story retelling by bilingual children with language impairments and typically developing controls,” International Journal of Language & Communication Disorders, vol. 49, no. 1, pp. 60–74, Aug. 2013.

[33] P. Escudero, G. Pino Escobar, M. Hernandez Gallego, C. Diskin-Holdaway & J. Hajek, “The Little Multilingual Minds corpus: Educators’ and children’s speech in heritage languages”, Workshop on community language corpora in Australia, Australian National University, Canberra, 2023.

[34] S.-Y. Yoon, L. Chen, and K. Zechner, “Predicting word accuracy for the automatic speech recognition of non-native speech,” Interspeech 2010, Sep. 2010.

Have Four-Year-Olds Mastered Vowel Reduction in English? An Acoustic Analysis of Bilingual and Monolingual Child Storytelling

Paola Escudero[1], *Weicong Li*[1] and *Chloé Diskin-Holdaway*[2]

[1] MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia [2] School of Languages and Linguistics, The University of Melbourne, Australia

paola.escudero@westernsydney.edu.au, weicong.li@westernsydney.edu.au,
chloe.diskinholdaway@unimelb.edu.au

Abstract

Recent studies show that late Spanish-English bilingual adults tend to under-reduce or over-reduce English vowels, with hardly any reduction in their Spanish vowels. We examined whether children acquiring Australian English and Spanish simultaneously show similar vowel reduction patterns. Bilingual four-year-olds' vowel acoustics (spectral and temporal) produced during story comprehension and retelling tasks were measured following previous methods used on child-directed speech. Unlike sequential bilingual adults, simultaneous bilingual four-year-olds mostly resembled monolingual peers in their English vowel reduction and acoustics. Findings are explained using the Second Language Linguistic Perception (L2LP) model, which uses input and language control to explain bilingual performance.

Index Terms: English vowel reduction, bilingual Spanish-English preschoolers, bilingual vowel production

1. Introduction

Vowel reduction is a well-known aspect of English phonology with a consensus on the fact that vowels in unstressed syllables are produced with less peripheral properties than stressed vowels [1-4]. What is less agreed upon is the specific properties of reduced vowels within and between varieties of English, with some researchers even referring to American English unstressed vowels as "targetless" [5] and highly dependent on position of the unstressed vowel in the word or sentence and the surrounding consonants [6]. Acoustic and articulatory studies of American English unstressed vowels have demonstrated that patterns of vowel reduction vary even within speakers [7].

There are fewer studies on vowel reduction in Australian English, but recent studies suggest that unstressed vowels are more likely to be reduced to the same central, mid vowel, in contrast to Southern British English where a higher and less central vowel is also used [8]. Adding to the variability in reduced vowels, [9] reports on formant values of Australian English vowels produced in spontaneous speech where many unstressed vowels are reduced to disappearance, e.g., "it's" produced as [ts] or [s] and the last syllable in "faces" and "places is produced as [s_z], with no vowel trace in any of these cases.

As mentioned by the authors in [9], English vowel reduction in natural connected speech may pose an even larger challenge not only for analysis but for an adult listener. The aim of the present study is to examine how this variability in vowel reduction affects language learning in bilingual populations, especially those for whom vowel reduction is not present in their language other than English.

Spanish-English bilinguals are a clear example of such bilinguals, with vowel reduction present in English but not in Spanish. The lack of vowel reduction in Spanish has been shown to contribute to high levels of foreign accentedness in the English speech of Spanish-English bilinguals [10]. Following [11] we define bilinguals as speakers of two languages who may have the same or different onsets of learning for the two languages. Bilinguals can be divided into simultaneous or sequential bilinguals, depending on the onset of language learning [11]. Within the sequential type, studies refer to early or late bilinguals, with the later commonly also referred to as second language (L2) speakers [e.g., 6].

Recent acoustic analyses have shown that late Spanish-English bilinguals tend to under-reduce or over-reduce some English vowels with respect to vowel formants and duration [12], while producing limited or no vowel reduction in their Spanish vowels [13]. High variability in the acoustic properties of English reduced vowels in both monolingual and bilingual adults has been highlighted in multiple studies, and explained by onset of learning, coarticulatory and lexical effects [6, 12]. For instance [14] found that early bilinguals differed less from monolinguals than late bilinguals, but that individual bilingual acoustic values varied more widely and displayed less group cohesiveness than monolingual values.

In this study, we focus on child rather than adult bilinguals. As mentioned in [15], studies comparing vowel acoustics in monolingual versus bilingual children are rare, with the authors mentioning a single study featuring Korean-English bilinguals [16]. Most studies on bilingual children's vowel production tend to use broad phonetic or phonological transcription [17, 18]. As mentioned above, vowel acoustic measurements are crucial to establish the extent of the variability of vowel reduction in both monolingual and bilingual children.

We could not find any previous acoustic study on English vowel reduction in children, but there are a few studies on bilingual children's vowel acoustics of English stressed vowels [see 19 for a review]. For instance, a recent study on Australian English measured the acoustics of /u/ produced by bilingual children with a variety of languages other than English (LOTEs) spoken at home. The authors found that bilingual children produced /u/ with a more retracted vowel than monolingual peers, with the bilingual pattern being correlated to the amount of use of the bilinguals' LOTE [20].

In the present study, we tested whether the development of vowel reduction in Spanish-English bilingual four-year-olds differs from that of age-matched monolingual peers, in line with previous studies with bilingual children and adults. In [15] it was reported that for American English stressed vowels, Spanish-English bilingual children behaved like late Spanish-English bilingual adults in that they did not exhibit the same level of acoustic vowel contrast distinction as age-matched monolingual children. However, the authors showed that language dominance played an important role, with English dominant bilingual children resembling monolingual children more closely than Spanish dominant bilinguals.

We therefore examined both stressed and unstressed vowels to compare our results of simultaneous Spanish-English bilingual children living in Australia to previous results on children and adults living in the US. As for most adult studies, acoustic analyses of bilingual children's productions have used recordings of isolated words, but we follow [9] in using connected speech recorded within an elicited story telling task to represent a more naturalistic setting for vowel reduction.

2. Method

2.1. Corpus and participants

The recordings used in the present study are part of a large corpus of child language speech in bilingual and monolingual children [21], as part of a longitudinal project on non-societal language maintenance and enhancement in Australia [22]. The data presented here is part of the first wave of data collection for the Spanish-English cohort which aimed at measuring proficiency in their two languages. Proficiency was measured with a battery of psycholinguistic tasks, two of which targeted their ability to understand and retell the story conveyed in a colourful and engaging audio-visual eBook [23].

Data collection took place via Zoom with an experimenter conducting a session where children participated in the tasks with the help of a parent for session set up (for details on the online protocol see [24]. The story comprehension and retelling component of the session lasted for approximately 10 minutes, including the presentation of the 12-page eBook.

Here we report on vowel acoustic data from a subset of 8 children who participated in the first session of the longitudinal project. Table 1 shows the number of children in the monolingual and bilingual groups together with important demographics information, including children's age and language input. This information was gathered using a Qualtrics survey form sent to parents together with the electronic consent form and study information sheet. Data collection was approved by the human ethics committee of the first two authors' university (ethics approval number: H11022).

Table 1. *Participants' demographic information*

	Spanish-English Bilinguals	English Monolinguals
n	5 (4 females)	3 (3 females)
Mean Age (range)	4.45 (4.1-4.9)	4.55 (4.2-5.3)
Mean % exposure English (range)	40.0 (10-90)	100
Mean % exposure Spanish (range)	60.0 (10-90)	0

2.2. Data processing

Audio recordings sampled at 16kHz or above were first extracted from Zoom sessions run on desktop/laptop computers. The suitability and reliability of using audio recordings via Zoom for linguistic studies have been studied by [25, 26]. To gain orthographical transcription at utterance level, the audio recordings were processed using Whisper developed by OpenAI [27]. Whisper is an automatic speech recognition system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. Resulting text files containing transcriptions and corresponding time information were converted to Praat [28] TextGrid files using MATLAB scripts.

Phonetically trained research assistants manually checked and corrected the transcriptions at utterance level. Using Whisper for ASR followed by manual review can substantially decrease human labour required for transcription, thus improving efficiency. Then the audio and transcription data were processed through WebMAUS [29] (Australian English model) to provide annotation at word and phoneme level. Boundaries and annotation were then manually checked by the research assistants who were instructed to use the automatic boundary unless it was obviously incorrect. We believe that the methodology of integrating the most recent technology such as Whisper in the workflow to speed up data processing could be widely applied to other linguistic research projects.

To annotate stressed and unstressed vowels, all the word tokens were first collated in a word list, which the research assistants went through and then manually marked the stressed/unstressed vowels. Then all the vowel tokens in the dataset were annotated as stressed/unstressed using the word list as a reference using a MATLAB script. Following [20], we annotated the Australian English vowel /ʉː/ followed by a lateral consonant as the allophone [ʊ], (e.g., SCHOOL), and the rest as [ʉː] (e.g., TWO).

2.3. Acoustic analysis

To examine vowel acoustics, duration and first and second format measurements were extracted in Praat. Formant measures were taken at 30 evenly distributed time points across the central portion of the vowel, starting from the 20% and ending with the 80% point [30-32]. For each vowel token, the series of 30-point vowel trajectories were smoothed using discrete cosine transform and averaged. Vowel tokens with fewer than 5 tokens were excluded. The final dataset included 1136 children's monophthong tokens: 718 by bilingual children and 418 by monolingual children. 714 were stressed syllables and 422 were unstressed syllables. Children's vowel acoustics were compared to those of the female adult speaker who read the eBook and conducted the Zoom session with the children, and who was a monolingual speaker of Australian English. She produced a total of 171 monophthong tokens.

As mentioned in the introduction, the main aim of the present study was to examine vowel reduction, specifically the amount of acoustic centralization in unstressed vowels, by both bilingual and monolingual children. To that end, we focused on the most peripheral vowels /ɪ, iː, æ, oː/, [ʉː], and [ʊ] and the most frequent unstressed vowels /ɪ, iː, æ, ə/, accounting for most child tokens and yielding approximately the same number of tokens for both vowel conditions (52% stressed = 371, 93% unstressed = 391). MATLAB and R were used for data visualization and statistical analyses on this subset of data, which we present below.

3. Results

Figure 1 shows ellipses for F1 and F2 values (mean and 2 standard deviations), together with individual tokens for stressed peripheral vowels and the most frequent unstressed vowel, namely /ə/. As expected of reduced vowels, /ə/ tokens have more centralised acoustic values than tokens of stressed peripheral vowels in child and adult productions. Multivariate (F1, F2 and duration) pairwise comparisons yielded a significant difference between peripheral vowels and /ə/ (all *p*s < 0.01), except for monolingual stressed [ʉː] and [ʊ] and bilingual stressed [ʊ] and unstressed /æ/.

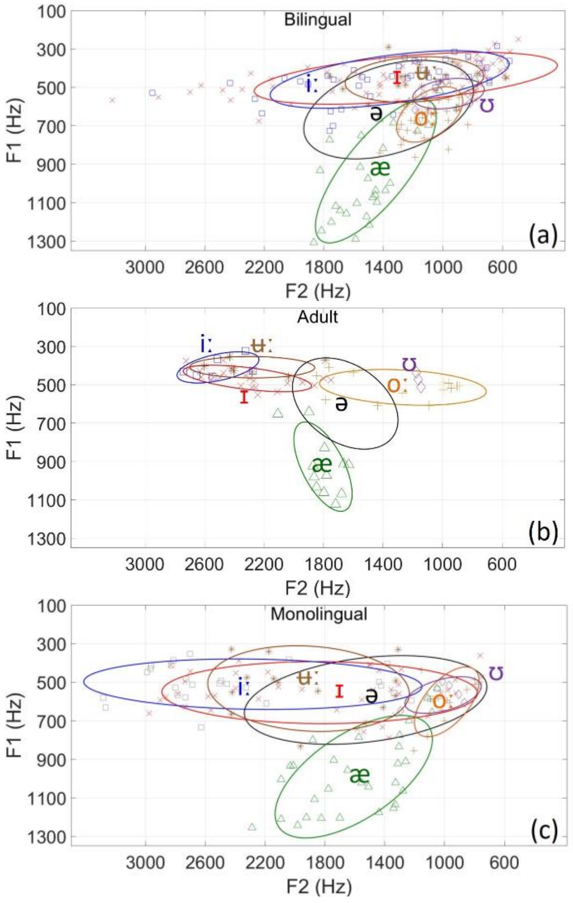


Figure 1: *Formant ellipses and data points for stressed peripheral vowels and* /ə/. *(a) bilingual children, (b) a monolingual adult, and (c) monolingual children.*

To further explore vowel reduction, we compared the acoustic values of the stressed and unstressed counterparts of /ɪ, iː, æ/. Separate linear mixed models for F1, F2 and duration with vowel, group, and stress condition as fixed effects and their interactions, and participants as random factors, revealed a significant interaction for group and vowel. Post-hoc pairwise comparisons showed that unstressed /æ/ had lower F1 values than its stressed counterpart for the adult (384 Hz lower) and bilingual children (268 Hz lower), bilingual unstressed /ɪ/ had lower F2 values (346 Hz lower), and monolingual unstressed /iː/ had lower F2 values (622 Hz lower) and longer duration (118 ms longer) (all *p*s < 0.05).

Children and adult productions seemed to differ only in some vowels and acoustic dimensions. Figures 2 and 3 show the largest differences in formant values with pairwise comparisons showing that the bilinguals and adult had lower F1 for unstressed /æ/ than monolinguals (at least 294 Hz lower), while bilinguals' F2 was lower than monolinguals and adult for stressed /ɪ, iː/ (at least 441 Hz lower) and [ʉː] (at least 546 Hz lower) and unstressed /ɪ/ (at least 727 Hz lower) (all *p*s < 0.01). Regarding duration, adult vowels (both stressed and unstressed) were 58–243 ms shorter than those of children, while bilingual and monolingual children had comparable vowel duration, with only stressed /iː/ 108 ms shorter in monolinguals than bilinguals (*p* = 0.03).

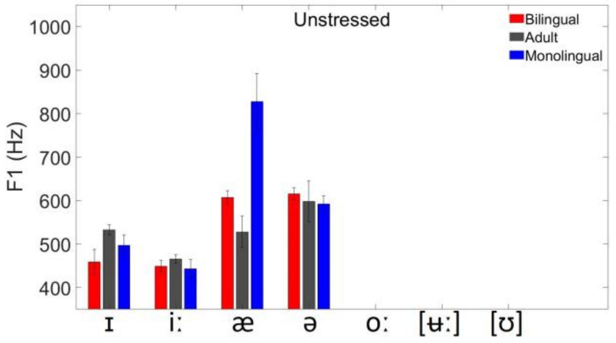


Figure 2: *F1 values (mean and standard error) of unstressed vowels.*

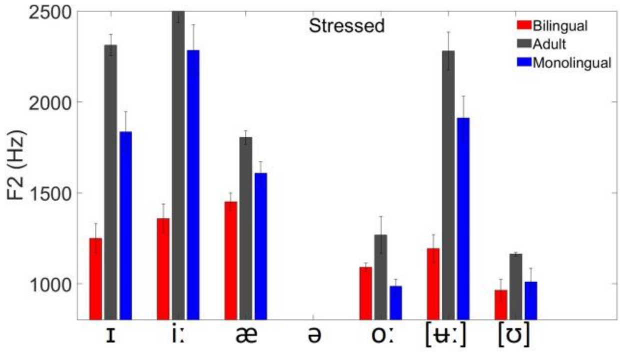


Figure 3: *F2 values (mean and standard error) of stressed vowels.*

4. Discussion and conclusion

As mentioned in the introduction, due to the lack of vowel reduction in Spanish, Spanish dominant bilingual children were expected to acquire English vowel reduction at a later stage in comparison to their English monolingual peers. This would be in line with previous studies with bilingual children and adults [15]. However, our findings demonstrate that 4-year-old children (even those with little exposure to English) have to a large extent mastered vowel reduction, as measured by the acoustic difference between their most frequent unstressed vowel /ə/ and their peripheral stressed vowels. Vowel reduction in children's productions was comparable to that of adult productions. Contrary to late bilinguals [12-14], bilingual children reduced unstressed /æ/ similarly to an adult, while monolingual children did not. This suggests that simultaneous bilingual children master vowel reduction more rapidly than monolingual peers.

The Second Language Linguistic Perception model (L2LP) [11, 33, 34, 35] proposes that input and the control of the different language modes are crucial variables for phonetic and phonological learning in bilinguals, predicting that amount of exposure and ability to inhibit the non-target language determine language learning [11]. In that respect, we also observed some qualitative differences: compared to monolinguals who showed a strong split between the prelateral and non-prelateral allophones, bilinguals showed much less divergence in F2 values when producing [ʉː] and [ʊ]. This finding may indicate the interplay between the bilinguals' two languages, as suggested in [20] for bilingual children in Australia with diverse LOTEs.

Influenced by Spanish, high back /ʉː/ vowels produced by caregivers could lead to bilingual children's realisation of /ʉː/ in two ways: as a more back [ʉː] or possibly as a less retracted [ʊ], resulting in further divergence from monolingual peer productions. Additionally, four-year-olds may not be able to inhibit or control their dominant language, i.e., Spanish, and exhibit English vowel productions that are influenced by their Spanish vowel productions. Ongoing research aims at understanding the interrelation between individual language exposure and vowel production in the two languages of a larger cohort of bilingual Spanish-English children. In the analysis of this larger corpus, we will also consider the effects of position in the word, vowel length and prosodic context [36].

5. Acknowledgments

This study was supported by three grants from the Australian Research Council awarded to the first author (CE140100041, FT160100514) and to the first and last author (LP210300631). We thank all children who participated in the study and their parents who provided consent for their participation, Deeahn Sako for data collection, Minh Anh Tran, Sharmin Saba, and Kristy Bailey for help with manual segmentation and adjustment of the vowel data.

6. References

[1] N. Chomsky, M. Halle, The Sound Pattern of English. New York: Harper & Row, 1968.

[2] K. Crosswhite, "Vowel reduction", In: B. Hayes, R. Kirchner, and D. Steriade (eds), Phonetically-based Phonology. Cambridge: CUP 191–231, 2004.

[3] E. Flemming and S. Johnson, "Rosa's roses: reduced vowels in American English", JIPA 37, 83–96, 2007.

[4] L. Burzio, "Phonology and phonetics of English stress and vowel reduction", Language Sciences 29(2-3), 154–176, 2007.

[5] Y. Kondo, "Targetless schwa: is that how we get the impression of stress-timing in English?" Proc. Edinburgh Linguistics Dept. Conf. 63–76, 1994

[6] E. Byers, M. Yavas, "Vowel reduction in word-final position by early and late Spanish-English bilinguals", PloS one, 12(4), e0175226, 2017.

[7] M. Proctor, C. Y. Lo, and S. Narayanan, "Articulation of English vowels in running speech: a real-time MRI study." In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences, International Phonetic Association, 2015.

[8] F. Cox and S. Palethorpe, "Rosa's roses – unstressed vowel merger in Australian English", Proc. 17th Australasian Int. Conf. on Speech Science and Technology (SST17). Australasian Speech Science and Technology Association (ASSTA), 89-92, 2018.

[9] G. Docherty, S. Gonzalez and P. Foulkes, "An acoustic study of the realisation of KIT in the conversational speech of young English speakers in Australia", Proceedings of the 20th International Congress of Phonetic Sciences, International Phonetic Association, 2023.

[10] J.E. Flege and O.S. Bohn, "An instrumental study of vowel reduction and stress placement in Spanish-accented English", Studies in Second Language Acquisition, vol.11, issue 1, pp. 35-62, 1989.

[11] P. Escudero P. and K. Yazawa, "The Second Language Linguistic Perception model (L2LP)," in Amengual, M. (Ed.). The Cambridge Handbook of Bilingual Phonetics and Phonology. Cambridge: CUP. Preprint. PsyArXiv, February 20, 2024. https://doi.org/10.31234/osf.io/qbx6z

[12] J.T. Conklin, O. Dmitrieva, Y.J. Jung and W. Zhai, "Acoustic characteristics of vowel reduction in advanced Spanish-English bilinguals", J. Acoust. Soc. Am. 148, no. 4_Supplement (2020): 2656-2657, 2020.

[13] J.T. Conklin, O. Dmitrieva, Y.J. Jung and W. Zhai, "Acoustic influence of L1 Spanish on L2 English vowel production", J. Acoust. Soc. Am. 150, no. 4_Supplement (2021): A42-A42, 2021.

[14] E. Byers, "Vowel reduction patterns of early Spanish-English bilinguals receiving continuous L1 and L2 input", Topics in Linguistics, 18(1), pp.17-31, 2017.

[15] S. van der Feest, G. Medina, E, Maryutina, I. Davidovich, T. Bloder, I. Barrière and V.L. Shafer, "Acoustic Correlates of Central Vowels in Russian-English and Spanish-English Bilingual Children", Proceedings of the 47th annual Boston University Conference on Language Development (BUCLD), 241-254. Somerville, MA: Cascadilla Press, 2023.

[16] S. Lee and G. Iverson, "Stop consonant productions of Korean–English bilingual children", Bilingualism: Language and Cognition, 15, 275–287, 2012.

[17] E. Jacewicz, R. Fox and A. Robert, "The effects of indexical and phonetic variation on vowel perception in typically developing 9- to 12-year-old children", Journal of Speech, Language, and Hearing Research, 57(2), 389-405, 2014.

[18] C.E. Gildersleeve-Neumann, E.S. Kester, B.L. Davis, E.D. Peña, "English speech sound development in preschool-aged children from bilingual English-Spanish environments. Lang Speech Hear Serv Sch. 39(3):314-28, 2008.

[19] J. Yang, "Vowel development in young Mandarin-English bilingual children." Phonetica. 22;78(3): 241-272, 2021.

[20] A. Gibson, F. Cox and J. Penny, "Acquiring allophony: GOOSE and SCHOOL vowels in the speech of Australian children", Proceedings of the 20th International Congress of Phonetic Sciences, International Phonetic Association, 2023.

[21] P. Escudero, G. Pino Escobar, M. Hernandez Gallego, Chloé, Diskin-Holdaway & J. Hajek, "The Little Multilingual Minds corpus: Educators' and children's speech in heritage languages", Workshop on community language corpora in Australia, Australian National University, Canberra, 2023.

[22] P. Escudero, G. Pino Escobar, C. Diskin-Holdaway, and J. Hajek. "Enhancing Heritage and Additional Language Learning in the Preschool Years: Longitudinal Implementation of the Little Multilingual Minds Program." OSF Preprints. September 16, 20204. doi:10.31219/osf.io/rvjcg.

[23] G. Pino Escobar and P. Escudero. "Vocabulary, Comprehension and Retelling in Multilingual Children: Age and Input Tell the Whole Story." OSF Preprints. September 16, 2024. doi:10.31219/osf.io/8e4nf.

[24] P. Escudero, G. Pino Escobar, C. G. Casey & K. Sommer, "Four-year-old's online versus face-to-face word learning via eBooks", *Frontiers in Psychology*, *12*, 450, 2021.

[25] C. Zhang, K. Jepson, G. Lohfink, & A. Arvaniti, A, "Speech data collection at a distance: Comparing the reliability of acoustic cues across homemade recordings", The Journal of the Acoustical Society of America, 148(4_Supplement), 2717-2717, 2020.

[26] C. Ge, Y. Xiong, & P. Mok, "How Reliable Are Phonetic Data Collected Remotely? Comparison of Recording Devices and Environments on Acoustic Measurements", Proceedings of Interspeech, pp. 3984-3988), 2021.

[27] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning (pp. 28492-28518). PMLR, July, 2013.

[28] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer" [Computer program]. Version 6.4.13, retrieved 10 June 2024 from http://www.praat.org/, 2024.

[29] T. Kisler, U. Reichel, and F. Schiel, “Multilingual processing of speech via web services”, Computer Speech & Language, vol. 45, pp. 326–347, 2017.

[30] D. Williams and P. Escudero, “A cross-dialectal acoustic comparison of vowels in Northern and Southern British English”, J. Acoust. Soc. Am. 136, 2751–2761, 2014.

[31] H. Sarvasy, J. Elvin, W. Li and P. Escudero, “An acoustic phonetic description of Nungon vowels”, J. Acoust. Soc. Am. 147, 2891–2900, 2020.

[32] H. Sarvasy, W. Li, J. Elvin and P. Escudero, “Vowel acoustics of Nungon child-directed speech, adult dyadic conversation, and foreigner-directed monologues”, Front. Psychol. 13:805447, 2022.

[33] K. Yazawa, J. Whang, M. Kondo and P. Escudero, “Language-dependent cue weighting: An investigation of perception modes in L2 learning.” Second Language Research, 36(4), 557-581, 2020.

[34] X. Liu, P. Escudero, “How bidialectalism affects non-native speech acquisition: Evidence from Shanghai and Mandarin Chinese”. Applied Psycholinguistics 44 (6), 969-990, 2023.

[35] K. Yazawa, J. Whang, M. Kondo and P. Escudero, “Feature-driven new sound category formation: computational implementation with the L2LP model and beyond”, Front. Lang. Sci. 2:1303511, 2023.

[36] I. Yuen, F. Cox, and K. Demuth, “Three-year-olds' production of Australian English phonemic vowel length as a function of prosodic context.” The Journal of the Acoustical Society of America 135(3), 1469-1479, 2014.

Table of Contents for this manuscript

Bilingual Preschoolers' Phonetic Variation Keeps Up with Monolingual Peers: The Case of Voiceless Plosives in Australian English

Chloé Diskin-Holdaway[1], *Weicong Li*[2] and *Paola Escudero*[2]

[1]School of Languages and Linguistics, The University of Melbourne, Australia
[2]MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia

chloe.diskinholdaway@unimelb.edu.au, weicong.li@westernsydney.edu.au, paola.escudero@westernsydney.edu.au

Abstract

We investigate phonetic variation in the English plosives /p t k/ of two groups of four-year-olds residing in Sydney, Australia: five Spanish-English bilinguals, and four Australian English monolinguals. Both groups are also compared with a monolingual adult. Nine categories of /p t k/ are identified across 901 tokens. No significant differences were found across groups, but /t/ exhibited the most variation, including variants such as a word-initial dentalized /t/, found only among the bilingual speakers. Results suggest that simultaneous bilingual children resemble monolinguals in their English plosives, with some minimal influence from their other language.

Index Terms: Australian English, Spanish, plosives, phonetic variation, simultaneous/early/late bilingualism.

1. Introduction

Phonetic variation in the voiceless plosives /p t k/ in Australian English (AusE) has been well-documented [1-4], with a particular interest in /t/, which has been shown to exhibit sociophonetic variation by dialect, region, age and gender [2]. There has been an emphasis on instances of glottalization, tapping (flapping), frication and deletion of these variables as compared to their canonical released variants, which are typically aspirated in word or syllable-initial position as [p^h], [t^h] and [k^h]. However, these consonants display considerable variability, with e.g. [2] listing 10 categories of /t/ among AusE (mainstream and Aboriginal) adult speakers.

Despite interest in the speech of adults, research on the phonetic development of these plosives among children has been limited, with exceptions including some work on British English and American English with children as young as three years of age [5-7]. A review of studies in [8] finds that children around five years of age are beginning to produce gender-based variation in the realisation of glottalized stops in BrE. Exceptions to this gap in work on AusE children include [9], which studied gender-specific variation of plosives among primary school-aged children (ages 5-12) in Yarrawonga, Victoria. An earlier study on connected speech processes among Brisbane adolescents also documented some variation in /t/, particularly its weakening in particular phonological environments, such as intervocalically (prone to tapping/flapping) and in phrase-final position (prone to unreleased variants) [10].

In [9], six categories of /p t k/ were examined across 18 children at three year levels (Prep, ages 5-6, Grade Three, ages 8-9 and Grade Six, ages 11-12). For /p/, which had the least number of tokens overall, most variants for both boys and girls were canonical released tokens, followed by an unreleased variant. The fricated variant appeared at a higher rate in the speech of boys than it did for girls. For /t/, which had the most tokens due to it occurring statistically more often in English (see also [10: 37], both the fricated (including the fully fricated and affricated realisations) and the pre-aspirated variants appeared in girls' speech at a higher rate than that of the boys, which is consistent with findings of these variants in adult speech [see e.g. 11: 7]. Glottalization of /t/ (including both glottal and laryngealized realisations) also occurred at a higher rate for girls than boys overall. The tapped variant appeared at around the same rate for both boys and girls. For /k/, most were canonical released tokens, with gender-specific variation patterning similarly to /p/. There was more frication of /p/ among boys compared to girls. Both /p/ and /k/ had more pre-aspiration among girls as compared to boys.

Despite [9] including three age groups, with children as young as five, age was not a variable that was considered in the analysis, so little is known about the effect of age on the sociophonetic patterning of /p t k/ in the Australian context. Furthermore, there are only two previous studies that have examined phonetic variation (in vowels) among bilingual children in Australia: [12] and [13], who studied children living in diverse areas of Sydney, coming from a variety of Language Other than English (LOTE) backgrounds, but not Spanish. There have been no previous studies looking at variation in plosives among children with a LOTE background in Australia. Spanish is the tenth most spoken language after English in Australia, spoken by 171,378 people or 0.7% of the population, [14], so there is a sizeable population of children growing up bilingually with both Spanish and English in the country.

We focus on a population of four-year-old simultaneous bilinguals who have been exposed to Spanish and English at home and in the community. Previous developmental research on 3-4 year-old Spanish-English bilinguals in the U.S. context [15] shows that those with higher exposure to English have a lower error rate for English consonants overall than those with equal exposure to English and Spanish. However, the phonemes /p t k/, which occur in both languages, did not present any particular difficulty unless they occurred in consonant clusters or word-finally, where deletion was common [15].

Other work in the U.S. context on Spanish-English bilingual adults shows that only early bilinguals with age of onset 12 or younger produce plosives with voice onset time (VOT) contrasts resembling those of English monolinguals [16] (see also [17]), but late bilinguals have different VOT values. Our research questions are thus:

-What are the patterns of /p t k/ production among Spanish-English bilingual children compared to AusE monolingual children?

-How do these patterns compare to that of an AusE monolingual adult experimenter?

-How is /p t k/ variation influenced by utterance environment?

Table of Contents for this manuscript

We predict that our bilingual group, all early simultaneous bilinguals, will have /p t k/ patterns resembling their monolingual peers, with some minimal differences expected for those with lower English exposure [15]. Since /p t k/ presents phonetic variability among somewhat older monolingual children [9], we investigate whether bilinguals also replicate this variability, or whether they have more canonical variants.

2. Method

2.1. Corpus and participants

The recordings used in the present study are part of a large corpus of child language speech in bilingual and monolingual children [18], as part of a longitudinal project on non-societal language maintenance and enhancement in Australia [19]. The data presented here is part of the first wave of data collection for the Spanish-English cohort which aimed at measuring proficiency in their two languages. Proficiency was measured with a battery of psycholinguistic tasks, two of which targeted their ability to understand and retell a story in English conveyed in a colourful and engaging audio-visual eBook [20].

Data collection took place via Zoom with an experimenter (the 'adult' in our analyses) conducting a session where children participated in the aforementioned two tasks with the help of a parent for session set up (for details on the online protocol see [21]). The story comprehension and retelling component of the session lasted for approximately 10 minutes, including the presentation of the 12-page eBook. For four of the participants and the adult experimenter, we also include their speech as it occurred in the lead-up to the two tasks, and in a subsequent task measuring their ability to reproduce nonce words. This was to bolster the number of tokens for greater statistical power and comparability with [9]. Therefore, our data includes connected speech (short sentences) and isolated words produced by the children during these psycholinguistic tasks. The adult experimenter's data is mostly connected speech containing directives and also some isolated words e.g. *Ready?*

	Bilinguals	Monolinguals
n	5 (4 females)	4 (4 females)
Mean Age (range)	4.40 (4.2-4.8)	4.56 (4.2-5.3)
Mean % exposure English (range)	40.0 (10-90)	100
Mean % exposure Spanish (range)	60.0 (10-90)	0

Table 1. *Participants' demographic information*

Here we report on 901 tokens of /p t k/ extracted from a subset of nine children. Table 1 shows the number of monolinguals and bilinguals and their gender, age and language input, as reported by the parents in a Qualtrics survey. Data collection was approved by the human ethics committee of Author 2 and 3's university (ethics approval number: H11022). We conduct group-level comparisons (monolinguals vs. bilinguals) as well as discussion of individual behaviour, which we consider to be important with this small sample size.

2.2. Data processing and labelling

Audio WAV files were first extracted from Zoom video recordings. OpenAI Whisper [22] was used for automatic speech recognition (ASR). The use of Whisper for ASR can substantially speed up the transcription process and reduce the required human input. Text transcriptions at utterance level, together with time information, were then converted into Praat TextGrid files using MATLAB scripts. Research assistants (RAs) manually reviewed and improved the transcriptions for accuracy. Then forced alignment using WebMAUS [23] with the AusE model was applied on the audio recordings and transcription to obtain annotation at both word and phoneme levels. The /p t k / annotations at phoneme level were extracted as a separate tier in Praat and coded (see section 2.3). For each token, the position in the word (initial, medial, or final), and preceding and following segment (vowel, consonant, or pause) were also extracted based on the annotation at word and phoneme level, resulting in a total of 3^3, i.e. 27 utterance environments. Data visualisation and statistical analysis were performed using MATLAB and R [24].

2.3. Phonetic categories for /p t k/

Nine codes were used to categorize tokens and were influenced by [2] and [9]. Each token was subject to auditory and visual analysis, with corresponding spectrogram inspected in Praat. Two RAs conducted coding, which was manually checked by Author 1. Coding was primarily top-down, with RAs relying on the codes, but discussing categorisations with Author 1. Inaudible tokens, or where the speaker missed the target, e.g., pronouncing *school* as [stul], were excluded on the basis of an absent /k/ and thus not conforming to adult categories.

1. **Affricate**: a "closure followed by an /s/-like release (not aspirated), no burst-like characteristics" [t^{s}] [2: 62].
2. **Deleted**: no visible or audible presence of the consonant. "Difficult to distinguish phonetically from unreleased /t/" [10: 40], and as such, used for cases where the consonant would typically be released, but is deleted to due to a developmental/L2-like 'error', e.g. *ask* pronounced as [as].
3. **Fricated**: identified acoustically by "high frequency energy and lack of stop closure and release" [9: 66]. Not the same as /s/, better described as a "lowered /t/" [2: 62].
4. **Glottalized**: includes two realisations: glottal /t/ ("no formant transitions and the presence of creaky phonation on either side of the stop closure") and laryngealized /t/ ("lack of stop closure or release and presence of fully laryngealized voicing throughout the segment") [9: 66].
5. **H**: only applies to /t/ when it is pronounced like a [h].
6. **Other**: voiced /k/, dental fricative /t/, dentalized /t/, etc.
7. **Release**: the 'canonical' aspirated variant. Characterised acoustically and visually by a "period of full closure followed by burst. No voicing apparent" [2: 62].
8. **Tap**: identified acoustically by "a short closure phase and a short period of voicing" [9: 66]. Only occurs intervocalically and only for /t/.
9. **Unreleased:** no visible or audible presence of the consonant (see also 'deleted' category above).

3. Results

There were 203 /k/ tokens (23% of the dataset), 178 /p/ tokens (19%) and 520 /t/ tokens (58%) (Table 2), mirroring ratios in [9], for /k/ (25%), /p/ (13%) and /t/ (62%). In terms of numbers, the affricate, glottalized, release and tap categories appeared to differ between the three groups in the present study, but a Mann-Whitney U-test (MWU) found no significance.

group	variable	Aff.	Del.	Fric.	Glott.	H	Other	Rel.	Tap	Unrel.	Total
Biling.	k	1	7	4	1	0	5	85	0	2	105
	p	0	0	1	0	0	1	70	0	2	74
	t	34	11	1	59	2	22	98	12	3	242
Monoling.	k	0	9	5	1	0	2	55	0	1	73
	p	0	12	1	0	0	0	76	0	2	91
	t	23	10	4	30	1	5	90	21	15	199
Adult	k	0	0	1	0	0	1	23	0	0	25
	p	0	0	0	0	0	0	13	0	0	13
	t	6	0	0	15	0	1	48	7	2	79

Table 2. *Number of variants for /p t k/ for ten speakers, broken down by group (biling.=bilingual, monling.=monolingual, adult). Aff. = affricate; Del. = deleted; Fric.= fricated; Glott.=glottalized; Rel. = released; Unrel.=unreleased.*

3.1. Canonical release

Canonical release is the most prevalent category, particularly for /p/ and /k/. The adult had high rates for /k/ (92%) and /p/ (100%) and somewhat lower for /t/ (60.76%). The bilinguals had higher rates for /k/ and /p/, and the monolinguals for /t/ (Table 3). The children's rates resemble the adult (highest for /p/, lowest for /t/), but with lower overall release rates, indicating more phonetic variation as compared to the adult.

	Bilinguals	Monolinguals
/k/	82.9 (69.7–93.3)	73.4 (66.7–82.8)
/p/	88.6 (66.7–100)	85.3 (71.4–100)
/t/	39.6 (22.7–52.4)	47.5 (39.7–65.5)

Table 3. *Mean percentage (and range) of release by group*

A linear mixed effects model (LMM) compared release percentage (dependent variable) by two language groups (bilingual and monolingual) and consonants (/p, t, k/), modelled as fixed effects (independent variables), and participant as a random effect, and showed that only release for /t/ was significantly lower than for /k/ (β = -43.3, p <.001). No significant interaction between language group and consonant was found. Post-hoc pairwise comparisons for /k/ and /p/ showed bilinguals had (not significantly) higher release percentages than monolinguals, but the opposite for /t/.

3.2. Utterance environment

	/k/			/p/			/t/		
	i	m	f	i	m	f	i	m	f
pp	0	0	0	0	0	0	1	0	0
pc	5	0	0	7	0	0	11	0	0
pv	14	0	0	32	0	0	98	0	0
cp	0	0	6	0	0	0	0	0	30
cc	5	6	4	6	3	3	4	16	25
cv	8	23	0	15	4	1	38	19	5
vp	0	0	45	0	0	25	0	0	82
vc	3	44	8	8	1	10	16	35	56
vv	8	15	9	29	32	2	28	39	17

Table 4. *Tokens of /p t k/ by word environment (i: initial, m: medial, f: final) and preceding/following segment (c: consonant, v: vowel, p: pause)*

The majority of /k/ tokens were in word-final position, preceded by a vowel and followed by a pause, as in *truck*, and in word-medial position preceded by a vowel and followed by a consonant, as in *lunchbox* (Table 4). Both were frequent words, occurring 51 and 33 times respectively. Most /t/ was in word-initial position preceded by a pause and followed by a vowel, as in *two*, and in word-final position preceded by a vowel and followed by a pause, as in *but*. These patterns have implications for /p t k/ variation, discussed further below.

3.3. Variation in /p/ and /k/

/p/ had the least phonetic variation (Fig. 1), being mostly canonically released and with many word-initial tokens (Table 4). Both bilingual and monolingual children had some fricated and unreleased variants for /p/. Two monolinguals had deleted variants, but only in the nonce word task (section 2.1) in words like *sep* and *ballop* where it was difficult to ascertain whether the target was deleted or not met at all.

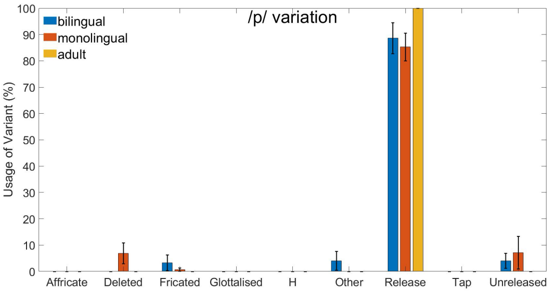


Figure 1: *Proportion of /p/ categories across three speaker groups (bilingual, monolingual, adult)*

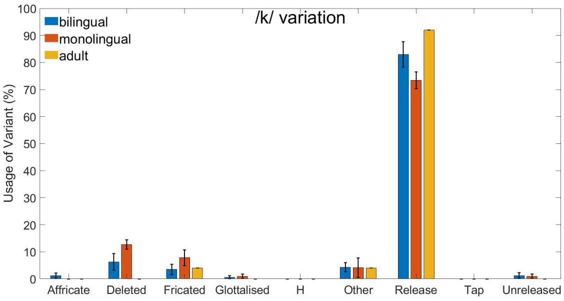


Figure 2: *Proportion of /k/ categories across three speaker groups (bilingual, monolingual, adult)*

/k/ had somewhat more variability than /p/ (Fig. 2), but no significant differences across groups and categories were found via MWU. Fricated /k/ was found across all groups, with some unreleased, deleted and glottalized tokens among the children. Monolinguals (12.33%) deleted somewhat more than bilinguals (6.67%), noting that many /k/ tokens were word-final and deleted in words such as *ask* and *desk*. All speakers had at least one /k/ in the 'other' category. All were voiced, except one palatalized /k/ for bilingual A018 in *take* and four ejectives for bilingual A017 in *okay*, *cake* (x2) and *truck*.

3.4. Variation in /t/

As well as being the most frequent of the three plosives, /t/ was also the most variable (Fig. 3), although no significant differences across groups and categories were found via MWU. The affricate category had somewhat higher proportions among the bilinguals (14.05%) and the monolinguals (11.56%) as compared to the adult (7.59%). In particular, the initial /t/ in *truck* was categorically affricated by the children, except one case of an ejective affricate by participant A017 and a case of deletion ([rʌk]), and epenthesis ([tʌra]) with a released /t/ by participant A016. The glottalized category was most prominent among the bilinguals (24.38%), followed by the adult (18.99%) and then the monolinguals (15.08%). However, the final /t/ context preceded by a vowel and followed by a pause had more glottalization by the monolinguals (53% in this context) than the bilinguals (39%) and was particularly prevalent in the phrase *don't know*. Tapped variants were also more prominent among the monolinguals (10.55%) as compared to the adult (8.86%) and the bilinguals (4.96%). The 'other' category was prominent among the bilinguals, but this was mainly due to A034, who had 28.95% 'other', including a dentalized /t/ in words like *Tom* [t̪ɔm] and *to* [t̪u] and a voiced variant of /t/ in *light*.

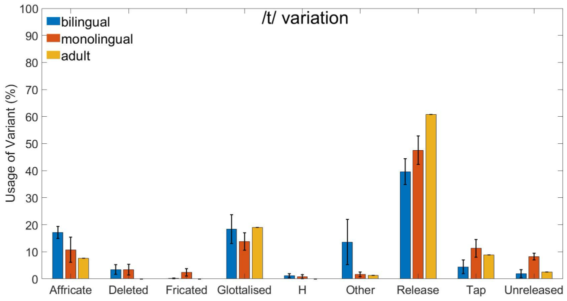


Figure 3: *Proportion of /t/ categories across three speaker groups (bilingual, monolingual, adult)*

4. Discussion and conclusion

We have presented phonetic variation in /p t k/ among a group of Spanish-English bilingual children as compared to AusE-speaking children and one adult. Overall, we find little quantitative evidence for major differences between the groups. Apart from some minor differences, the children appear to have mastered adult-like phonetic variation in plosives, and bilingual children match monolingual peers. We note that the adult had less variable /p t k/ overall as compared to the children and to other studies on AusE-speaking adults [e.g., 2]. As experimenter, they were likely engaging in careful speech to come across as clear and articulate. Child-directed speech has also been found to enhance discriminability of stop consonants via exaggeration of VOT contrasts [25].

/t/ presented the most phonetic variability across all groups, consistent with other work on /t/ in AusE [e.g. 2]. However, the 'other' category for /t/ was notable among participants such as A034, who also had one of the highest rates of Spanish listening (90% of the time) and Spanish speaking (95%; aggregate exposure scores in Table 1). These variants, particularly a dentalized /t/ in syllable-onset position, can be traced to direct influence from the Spanish denti-alveolar consonant. Qualitative observations among other participants with relatively high Spanish listening/speaking (A019 – 50%; 50%; A016 – 86%; 94%) suggest other influences of Spanish, such as coda /s/ deletion and epenthesis in consonant clusters (also found in [15] in the U.S. context).

We note some further developmental observations: the word *lunchbox* presented some difficulty, with several bilingual children, e.g. A019 and A016, variably deleting the final /s/, pronouncing the word as [lʌnʃbɔk]. Participant A016 had five final /s/ deletions across their tokens of /p t k/. Participants A016 and A034 also introduced vowel epenthesis in the consonant cluster ([lʌnʃəbɔks]) in some *lunchbox* tokens, which could also be an influence of the word boundary (*lunch* and *box*). While epenthesis is common in children's speech, it is worth noting that this phenomenon only occurred among the bilingual children in this sample. However, the monolingual children also exhibited some likely developmental differences in their deletion of the final /p/ in nonce words like *sep* and *ballop*, at similar frequencies to the 'other' category for /t/ among the bilinguals.

In comparison with plosives in AusE monolingual children [9], we find lower rates of unreleased variants and somewhat higher rates of released variants, particularly for /p/ and /t/. However, this could be explained by the fact that [9] included more connected speech (elicited through "spontaneous conversation and interactive games and activities") than our dataset, whereas we had children speaking both in short sentences and using standalone words, usually in response to a question as part of a task led by the experimenter. Despite this, [9] had similar proportions of token numbers of /p/:/t/:/k/ as in our study. Other differences include higher rates of /t/ glottalization in our sample: as high as 24% for the bilinguals versus a high of 10% in [9]. Conversely, our participants have lower rates of /t/ and /k/ frication than in [9] and compared to adult AusE speakers [1, 26]. Our rates of tapping are also lower than in [9] (rates close to 20%), but our monolinguals and adult (8-10%) have rates closer than the bilinguals (4.96%).

As regards utterance environment, the word-final or coda environment for /t/ presents itself as a site of notable variation, particularly favourable to phenomena such as glottalization. Future work could consider the interaction between voice quality and the coda /t/ environment with this sample [27]. Our preliminary observations show that the children exhibit creaky voice, breathy voice and whispery voice. We also plan to examine VOT and voice termination time (VTT) [see 28] in the future, and to investigate the effects of word frequency and individual differences. Previous work has found adult early simultaneous Spanish-English bilinguals to have mastered VOT in English (as compared to late bilinguals with "compromise" VOT values [17]; see also [16]). We would like to provide a first indication that simultaneous bilingual children acquire the ability to inhibit the non-target language from an early age and more effectively than sequential bilinguals, as proposed by the L2LP model [29, 30].

5. Acknowledgements

Supported by Australian Research Council grants to Author 3 (CE140100041, FT160100514) and Author 1 and Author 3 (LP210300631). With thanks to participating children and parents, Deeahn Sako for data collection, and Minh Anh Tran and Linh Tran for coding assistance.

6. References

[1] D. Loakes and K. McDougall, "Individual Variation in the Frication of Voiceless Plosives in Australian English: A Study of Twins' Speech," *Australian Journal of Linguistics,* vol. 30, no. 2, pp. 155-181, 2010.

[2] D. Loakes, K. McDougall, and A. Gregory, "Variation in /t/ in Aboriginal and Mainstream Australian Englishes," in *Australasian Speech Science and Technology Conference*, Canberra, R. Billington, Ed., 2022: ASSTA, pp. 61-65.

[3] L. Tollfree, "Variation and change in Australian consonants: reduction of /t/," in *Varieties of English Around the World: English in Australia*, D. Blair and P. Collins Eds. Amsterdam: John Benjamins, 2001, pp. 45-67.

[4] B. Horvath, *Variation in Australian English: The Sociolects of Sydney*. Cambridge; New York: Cambridge University Press, 1985.

[5] P. Foulkes, G. J. Docherty, and D. J. L. Watt, "The emergence of structured variation," *University of Pennsylvania Working Papers in Linguistics,* vol. 7, no. 3, pp. 67-84, 2001.

[6] G. Docherty, P. Foulkes, B. Dodd, and L. Milroy, "The emergence of structured variation in the speech of Tyneside infants," Final report to the United Kingdom Economic and Social Research Council, grant R000 237417, 2002.

[7] J. Roberts, "As old becomes new: Glottalization in Vermont," *American Speech,* vol. 81, no. 3, pp. 227-249, 2006.

[8] J. Milroy, L. Milroy, S. Hartley, and D. Walshaw, "Glottal stops and Tyneside glottalization: Competing patterns of variation and change in British English," *Language Variation and Change,* vol. 6, no. 3, pp. 327-357, 1994.

[9] C. Tait and M. Tabain, "Patterns of gender variation in the speech of primary school-aged children in Australian English: the case of /p t k/," in *Australasian Speech Science and Technology Conference*, Parramatta, 2016: ASSTA, pp. 65-68.

[10] J. C. L. Ingram, "Connected speech processes in Australian English," *Australian Journal of Linguistics,* vol. 9, no. 1, pp. 21-49, 1989.

[11] D. Loakes, K. McDougall, J. Clothier, J. Hajek, and J. Fletcher, "Sociophonetic variability of post-vocalic /t/ in Aboriginal and mainstream Australian English," in *Australasian Speech Science and Technology Conference*, Sydney, 2018: ASSTA, pp. 5-8.

[12] J. Penney, F. Cox, and A. Gibson, "Hiatus resolution and linguistic diversity in Australian English," *Phonetica,* vol. 81, no. 2, pp. 119-152, 2024.

[13] A. Gibson, F. Cox, and J. Penney, "Acquiring allophony: GOOSE and SCHOOL vowels in the speech of Australian children," in *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czechia, R. Skarnitzl and J. Volín, Eds., 2023: Guarant International, pp. 3750-3754.

[14] Australian Bureau of Statistics (ABS), "Census of Population and Housing," 2021. Accessed: 18 June 2024. [Online]. Available: https://profile.id.com.au/australia/language

[15] C. E. Gildersleeve-Neumann, E. S. Kester, B. L. Davis, and E. D. Peña, "English speech sound development in preschool-aged children from bilingual English-Spanish environments," *Language, Speech and Hearing Services in Schools,* vol. 39, no. 3, pp. 314-328, 2008.

[16] D. F. Thornburgh and J. H. Ryalls, "Voice onset time in Spanish-English bilinguals: Early versus late learners of English," *Journal of Communication Disorders,* vol. 31, no. 3, pp. 215-229, 1998.

[17] J. E. Flege, "Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language," *Journal of the Acoustical Society of America,* vol. 89, no. 1, pp. 395-411, 1991.

[18] P. Escudero, G. Pino Escobar, M. Hernandez Gallego, and C. Diskin-Holdaway, "The Little Multilingual Minds corpus: Educators' and children's speech in heritage languages," presented at the Workshop on community language corpora in Australia, Canberra, 2023.

[19] P. Escudero, G. Pino Escobar, C. Diskin-Holdaway, and J. Hajek, "Enhancing Heritage and Additional Language Learning in the Preschool Years: Longitudinal Implementation of the Little Multilingual Minds Program," *OSF Preprints,* September 16 2024, doi: doi.org/10.31219/osf.io/rvjcg.

[20] G. Pino Escobar and P. Escudero, "Vocabulary, Comprehension and Retelling in Multilingual Children: Age and Input Tell the Whole Story," *OSF Preprints,* September 16 2024, doi: doi.org/10.31219/osf.io/8e4nf.

[21] P. Escudero, G. Pino Escobar, C. G. Casey, and K. Sommer, "Four-Year-Old's Online Versus Face-to-Face Word Learning via eBooks," *Frontiers in Psychology,* vol. 12, 2021, Art no. 610975.

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023: PMLR, pp. 28492-28518.

[23] T. Kisler, U. D. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language,* vol. 45, pp. 326–347, 2017.

[24] *R: A Language and Environment for Statistical Computing*. (2024). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org

[25] B. McMurray, K. A. Kovack-Lesh, D. Goodwin, and W. McEchron, "Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence?," *Cognition,* vol. 129, no. 2, pp. 362-378, 2013.

[26] M. J. Jones and K. McDougall, "A comparative acoustic study of Australian English fricated /t/: assessing the Irish (English) link," in *Australasian Speech Science and Technology Conference*, Auckland, P. Warren and C. I. Watson, Eds., 2006: ASSTA.

[27] J. Penney, F. Cox, and A. Szakay, "Glottalisation, coda voicing, and phrase position in Australian English," *Journal of the Acoustical Society of America,* vol. 148, 2020, Art no. 3232.

[28] R. Mailhammer, S. Sherwood, and H. Stoakes, "The inconspicuous substratum: Indigenous Australian languages and the phonetics of stop contrasts in English on Croker Island," *English World-Wide,* vol. 41, no. 2, pp. 162-192, 2020.

[29] P. Escudero, *Linguistic Perception and Second Language Acquisition: Explaining the attainment of optimal phonological categorization*. Utrecht University: LOT Dissertation Series 113, 2005.

[30] P. Escudero and K. Yazawa, "The Second Language Linguistic Perception Model (L2LP)," in *The Cambridge Handbook of Bilingual Phonetics and Phonology*, M. Amengual Ed. Cambridge: CUP, in press.

English Lexical Productivity and Diversity in Spanish-English Bilingual Children in Australia

Milena Hernández Gallego[1], *Gloria Pino Escobar*[1], *Weicong Li*[1] and *Paola Escudero*[1]

[1] MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia

m.hernandezgallego@westernsydney.edu.au, g.pinoescobar@westernsydney.edu.au, weicong.li@westernsydney.edu.au, paola.escudero@westernsydney.edu.au

Abstract

In English-speaking countries such as Australia, the English proficiency of bilingual children, who also speak a heritage language (HL), has persistently been a matter of concern. Research indicates that bilingual children may initially show weaker vocabulary skills than monolingual peers. However, increasing evidence indicates that this gap is not due to bilingualism, but to individual learning experiences. We aimed at contributing to this debate by comparing the lexical productivity of 4 monolingual (English) and 5 simultaneous bilingual (Spanish-English) preschool children. Bayesian analysis revealed similar performance for the two groups, suggesting that bilingualism does not hinder English lexical development. Bilinguals can therefore achieve similar lexical productive and density outcomes as monolingual children before starting school. Implications for models of bilingual language acquisition will be discussed.

Index Terms: Bilingual children, lexical productivity, lexical density, Spanish-English bilinguals, simultaneous bilinguals.

1. Introduction

Acquiring a heritage language (HL) alongside a societal language is common around the world, including in Australia where about 30% of households speak an HL other than English [1]. Children from these households naturally become bilingual. There are concerns about whether these HL children develop an adequate expressive vocabulary [2], which is a crucial skill that predicts later vocabulary growth [3], school readiness, reading skills and academic success for monolingual and bilingual children [4]. Specifically, studies have shown that a child's early vocabulary size in both English and their HL predicts school readiness [5], as well as later reading skills [6].

When tested in one of their languages, bilingual children typically score lower on vocabulary tests than monolingual peers [7,8]. There is, however, some evidence showing that bilinguals catch up with monolingual norms starting at around 8 years of age [9] and by the mid-school years [10]. Additionally, some studies [11], using the American English and Spanish version of the CDI [12] have looked at bilingual children's lexical production and diversity in each language compared to monolinguals and have found comparable results. Other studies [13] further controlled for SES and found that bilingual children from middle-class families showed comparable receptive and expressive vocabulary to monolinguals. However, the studies mentioned above suggest that current findings are contradictory, without a clear indication of whether bilingualism yields substantial and important differences in expressive vocabulary when children transition to primary school.

Traditional views suggest that differences stem from bilinguals receiving limited exposure in each of their languages, which can result their achievement of developmental milestones at a different pace (usually later) than monolinguals of either language [14]. Bilinguals must split their exposure between the two languages, and some accounts hold that lexical representations in bilingualism are therefore weaker and have higher activation thresholds [15].

Indeed, bilingualism takes place along a continuum that implies more individual variability than in monolingual development [16, 17]. Within this continuum, studies of early bilingual development have consistently found that the relative amount of speech addressed to children is a strong predictor of children's skill development in both languages [18,19], including a larger vocabulary size [20]. For instance, the simultaneous bilingual children in [14] had higher within-group variation than monolinguals, but those with predominant English exposure compared well with monolingual peers.

Regarding the development of expressive vocabulary in context, monolingual children start to retell simple stories with adult guidance at 3 years [21], and at the age of 4, their narratives show basic story structure and a better understanding of the main events [22]. Due to the high variability in bilingual children linguistic experience [16, 17], it has not been confirmed whether these age-related milestones also apply to them [8].

Most studies looking at children's expressive vocabulary have focussed on HL children younger than three years old [8]. Many of the studies with children above 3 have been conducted within programs such as the US Head Start for children from low-income families [6,23]. However, it is still unknown how their results translate to the Australian context, that contrary to other English-speaking countries like the U.S., does not have a clear dominant HL [24]. Additionally, many studies on bilingual children lexical development have used parental reports or have used productive vocabulary tests [11, 13, 14]. Few studies have investigated word retrieval within narratives, as this discourse skill requires a different complexity in retrieval, as words must be related to a topic [25] rather than just producing labels for images.

Here we measure children's story retelling productions at the microstructure level to understand how bilingualism affects the lexical development required for more sophisticated oral discourse skills. The microstructure of a narrative includes lexical, morphosyntactic, and syntactic knowledge that is frequently operationalized in terms of story length or lexical productivity (e.g., total number of words and utterances) and lexical diversity (e.g. number of different words) [26]. Numerous studies have demonstrated variations in microstructure between bilingual children's two languages (e.g., Spanish and English [25]).

To measure children’s lexical productivity, we used the number of words and number of utterances (TNW and NU) because one measures word frequency and the other longer utterances involved in dialogue and communication, which can be used as indicators language development [3, 27]. We also used the number of different words (NDW) as an indicator of vocabulary breadth and richness, to quantify the lexical diversity of the children’s retelling productions [28]. Previous studies have used these measures to investigate lexical productivity and diversity of bilingual children, how they differ across languages and the impact of exposure or dominance over them [25, 29], agreeing that a higher level of language proficiency typically involves greater lexical productivity and diversity in the children’s output.

In the present study, we aim to bridge the gap in our understanding of how chronological age, linguistic input, and lexical productivity and diversity interrelate in Spanish-English bilingual children. Specifically, we used NU, TNW & NDW to compare bilingual to monolingual lexical development for oral discourse skills.

2. Method

2.1 Corpus and participants

The recordings used in the present study are part of a large database of child language speech in bilingual and monolingual children, as part of a longitudinal project on HL language maintenance and enhancement in Australia [30, 31]. The data presented here is part of the first wave of data collection that included Spanish-English bilingual children’ linguistic and cognitive assessments in their two languages. Linguistic proficiency was measured with a battery of psycholinguistic tasks, two of which targeted their ability to understand and retell the story conveyed in a colourful and engaging audio-visual eBook. Data collection took place via Zoom with an experimenter conducting a session where children participated in the tasks with the help of a parent for session set up [for details on the online testing protocol see [18, 32]. The story comprehension and retelling component of the session lasted for approximately 10 minutes, including the presentation of the 12-page eBook.

Here we report on speech data from a subset of 9 children who participated in the first session of the longitudinal project. Table 1 shows the number of children in the monolingual and bilingual groups together with demographics information, including children’s age and language input. This information was gathered using a Qualtrics survey form sent to parents together with the electronic consent form and study information sheet. Data collection was approved by the Western Sydney University human ethics committee (ethics approval number: H11022).

Table 1: *Participants demographic information*

	Spanish- English bilingual	English Monolingual
n	5	4
Females (n)	3	3
Mean Age (range)	4.44 (4.0-4.9)	4.30 (4.1-5.2)
Mean English exposure % (range)	40.00 (14-75)	100
Mean Spanish exposure % (range)	59.20 (25-86)	0
Median Principal Carer Relati	Mother (n=5)	Mother (n=3)
Median Principal Carer Education	University degree (n=5)	University degree (n=3)

As shown in the table all bilingual participants received a minimum of 40% of English input and 59% of Spanish input. From the 5 bilingual children, 3 reported Spanish to be their main language. Socioeconomic status (SES) was accounted for using parents/carers education as a proxy, with all the participants principal carer having completed a university degree.

2.2 Data processing

Initially, audio WAV files were extracted from Zoom video recordings. These files underwent orthographical transcription at the utterance level using OpenAI's Whisper [33], an automatic speech recognition (ASR) system that is trained on approximately 680,000 hours of diverse, multilingual data sourced from the internet. The text transcriptions and time information were then converted into Praat TextGrid files using MATLAB scripts. The first author manually reviewed and refined these transcriptions for accuracy. Compared with transcribing audio manually, the use of Whisper for ASR followed by manual review substantially decreases human resource required for transcription and improves efficiency. Further, forced alignment can be applied after ASR to provide annotation at phoneme level. We believe that employing such methodology would speed up various types of linguistic research projects.

2.3 Statistical analysis

To examine lexical density and diversity, the number of utterances, words and different words from the used speech were annotated into a dataset. The dataset included a total of 485 children’s utterances, 1570 words and 539 unique words. JASP [34] was used for data visualization and statistical analyses. We conducted two Bayesian independent sample T-Tests to investigate the probability of the groups (Spanish-English bilingual vs English monolingual) differing in number of utterances (NU) and number of different words (NDW), and a Bayesian Mann-Whitney test to explore what the probability of the groups differing in number of words (TNW) was, as the data for the TNW variable was not normally distributed.

We chose a Bayesian t-test over its frequentist analogue because Bayesian statistics offer a more intuitive and comprehensive analysis in estimating the probability of an effect. Specifically, Bayesian t-tests estimate the full distribution of credible values for parameters and use Bayes factors to quantify evidence for or against hypotheses. This approach allows for more nuanced conclusions than the binary reject/accept decision based on p-values [35].

3. Results

Figures 1-3 show descriptive intervals for the mean of each variable (NU, TNW, and NDW). The mean of the NU (monolingual M=47.8, SD=9.8, bilingual M=53.3, SD=13.9) and the TNW (monolingual M=152.6, SD=44.8, bilingual M=167.5, 85.5) is higher for the bilingual group, and higher for the monolingual group in the NDW variable (monolingual M=58.6, SD=18.7, bilingual M=53, SD=32.3). However, it can be observed that the standard deviation (SD) is higher for the bilingual group for the three variables.

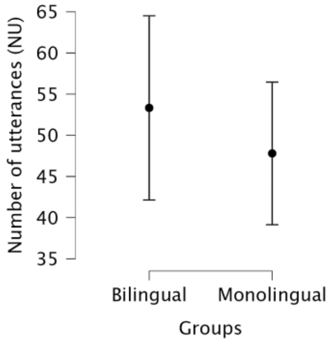


Figure 1: *Mean and standard deviation of the number of utterances for each group.*

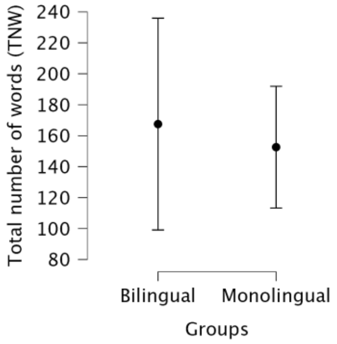


Figure 2: *Mean and standard deviation of the total number of words for each group.*

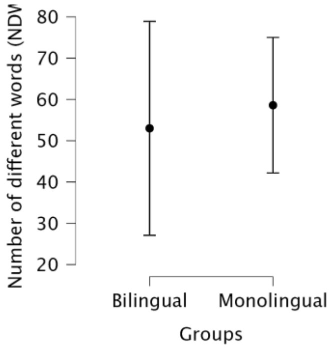


Figure 3: *Mean and standard deviation of the number of different words for each group.*

In Bayesian t-tests the null hypothesis (H_0) posits that there is no difference between the groups, while the alternative hypothesis (H_1) suggests that difference exists. The likelihood of the data under each hypothesis is calculated to provide Bayes factors, which indicate the strength of evidence for one hypothesis over the other [35]. Bayesian t-tests were used to compare the distributions of NU and NDW between the two groups, estimating the full range of credible values for the differences.

For the NU variable, the results showed that the data are 2.269 times more likely under the alternative hypothesis (H_1) than under the null hypothesis (H_0). According to the American Statistical Association (ASA) guidelines [36], a Bayes factor (BF_{10}) between 1 and 3 is considered to provide anecdotal evidence for H_1. This suggests weak evidence supporting a difference in the number of utterances between Spanish-English bilinguals and monolinguals.

Regarding the TNW, Bayesian Mann-Whitney test, a non-parametric alternative, compared the rank distributions of TNW between the two groups. Results indicated that the data are 0.594 times more likely under H_1 than under H_0, which translates to a Bayes factor of approximately 1.68 for H_0 (1/0.594) [36]. This indicates anecdotal evidence in favour of H_0, suggesting that there is little to no difference in the total number of words used between the two groups. Finally, regarding the NDW variable, the data are 0.513 times more likely under the H_1 than under H_0, corresponding to a Bayes factor of approximately 1.95 for H_0 (1/0.513). Like the TNW result, this provides anecdotal evidence favouring the null hypothesis, implying that there might not be meaningful difference in the number of unique words used between Spanish-English bilinguals and monolinguals.

4. Discussion and conclusion

The present study aimed at contributing to our understanding of how bilingualism influences lexical acquisition required for sophisticated oral discourse skills such as retelling a story. Bilingual and monolingual four-year-olds' lexical productivity and diversity when retelling an English story was compared by measure of their number of utterances (NU), total number of words (TNW) and number of different words (NDW). For all three measures, SD in the bilingual group was higher than that of the monolingual group, in line with previous studies [16,17] and the fact that bilingualism implies more variation than monolingualism. Despite the larger variation for bilinguals, we found anecdotal but not strong evidence suggesting a group difference for NU, while the evidence slightly favoured the null hypothesis for TNW and NDW, indicating no significant difference between the groups.

These findings suggest that bilinguals' limited exposure to each of their two languages compared to monolinguals does not necessarily lead to lower lexical productivity and diversity [20]. This is particularly the case of children growing up in high SES households [13], as is the case for the bilingual children included in the present study (Table 1). Our limited sample size, however, and our focus on measuring quantity rather than quality may have resulted in a less pronounced difference between bilingual and monolingual children.

Ongoing research includes a larger sample of children in both the monolingual and bilingual groups with more detailed information of their demographic background such as how SES and input exposure affect performance. We will consider a higher number of utterances by analysing the whole recording per child (45-60 minutes rather than the 10 minutes reported on in the present study), as well as an analysis of specific

Table of Contents for this manuscript

grammatical categories (e.g., nouns, verbs, adjectives) and grammatical errors, which will improve our understanding of the effects of bilingualism on lexical acquisition for sophisticated oral discourse skills.

Further research should also aim at examining the effect of quantity and quality of the bilingual input, and whether enhancing HL input through HL enhancement and maintenance projects such as [18, 30, 31] yields positive results for bilinguals. Additionally, more research on establishing whether monolingual story retelling age-related milestones apply to bilinguals would be important to finally shift the current view of bilingualism as a deficit and toward embracing it as a distinct linguistic capability [16].

Despite the limitations acknowledged above, we believe the current findings support the benefits of bilingual education and reassure parents that maintaining a heritage language alongside English will not hinder their children's language development, as was proposed in [18]. Additionally, this study and future research, adopting an individual differences approach could have implications such as enhancing the accuracy of information accessible to clinicians regarding bilingual development (e.g., to avoid overlooking a bilingual child who exhibits delays in both languages, because clinicians incorrectly anticipate that delays in both languages are typical).

5. Acknowledgements

This study was supported Australian Research Council grants awarded to the fourth author (CE140100041, FT160100514, LP210300631). We thank all children who participated in the study and their parents who provided consent for their participation. Special thanks to Anthony Serrano for his assistance with preliminary data pre-processing and analysis within his summer research internship under the supervision of the second and fourth authors, funded by the Western Sydney University Summer Scholarship program. We also thank Deeahn Sako for her help with data collection.

6. References

[1] Australian Bureau of Statistics. “Cultural diversity: Census information on country of birth, year of arrival, ancestry, language and religion”, www.abs.gov.au, Apr. 07, 2022. https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release

[2] M. J. Kieffer, “Early oral language and later reading development in Spanish-speaking English language learners: Evidence from a nine-year longitudinal study,” Journal of Applied Developmental Psychology, 33 (3): 146–157, 2012.

[3] L. Fenson et al., “Variability in early communicative development,” Monographs of the Society for Research in Child Development, 59 (5): i–185, 1994.

[4] A. Biemiller, “Vocabulary: needed if more children are to read well,” Reading Psychology, vol. 24, no. 3–4, pp. 323–335, 2003.

[5] C. S. Tamis-LeMonda et al., “Children’s Vocabulary Growth in English and Spanish Across Early Development and Associations With School Readiness Skills,” Developmental Neuropsychology, 39 (2): 69–87, 2014.

[6] M. D. Davison, C. Hammer, and F. R. Lawrence, “Associations between preschool language and first grade reading outcomes in bilingual children,” Journal of Communication Disorders, 44 (4): 444–458, 2011.

[7] E. Hoff, R. Rumiche, A. Burridge, K.M. Ribot, and S.N. Welsh, “Expressive vocabulary development in children from bilingual and monolingual homes: A longitudinal study from two to four years”. Early childhood research quarterly, 29(4): 433-444, 2014.

[8] P. Uccelli, and M.M. Páez, “Narrative and vocabulary development of bilingual children from kindergarten to first grade: Developmental changes and associations among English and Spanish skills, 2007.

[9] G. Jia, J. Chen, H. Kim, P. S., Chan, and C. Jeung, “Bilingual lexical skills of school-age children with Chinese and Korean heritage languages in the United States”, International Journal of Behavioral Development, 38(4), 350-358, 2014.

[10] V. C. M. Gathercole and E. M. Thomas, “Bilingual first-language development: Dominant language takeover, threatened minority language take-up”, Bilingualism: language and cognition, 12(2), 213-237, 2009.

[11] B.Z. Pearson, S.C., Fernández, S. C., & D.K. Oller, “Lexical development in bilingual infants and toddlers: Comparison to monolingual norms.” Language learning, 43(1), 93-120, 1993.

[12] D. Jackson-Maldonado, V. A. Marchman, and L. C. H. Fernald, “Short-form versions of the Spanish MacArthur–Bates Communicative Development Inventories,” Applied Psycholinguistics, 34(4): 837–868, 2012.

[13] A. De Houwer, M.H. Bornstein, and D. L. Putnick, “A bilingual–monolingual comparison of young children’s vocabulary size: Evidence from comprehension and production,” Applied Psycholinguistics, vol. 35, no. 06, pp. 1189–1211, 2013.

[14] E. Hoff, C. Core, S. Place, R. Rumiche, M. Senor, and M. Parra, “Dual language exposure and early bilingual development,” Journal of Child Language, 39 (1): 1–27, 2011.

[15] T. H. Gollan, R. I. Montoya, C. Cera, and T. C. Sandoval, “More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis☆,” Journal of Memory and Language, 58(3): 787–814, 2008.

[16] J. Paradis, “Sources of individual differences in the dual language development of heritage bilinguals,” Journal of Child Language, 50 (4), 1–25, 2023.

[17] J. Rothman, F. Bayram, V. DeLuca, Jorge González Alonso, M. Kubota, and Eloi Puig-Mayenco, “Chapter 3. Defining bilingualism as a continuum,” Studies in bilingualism, pp. 38–67, 2023.

[18] G. Pino Escobar and P. Escudero. “Vocabulary, Comprehension and Retelling in Multilingual Children: Age and Input Tell the Whole Story.” OSF Preprints. September 16, 2024. doi:10.31219/osf.io/8e4nf.

[19] E. Thordardottir, “The relationship between bilingual exposure and vocabulary development,” International Journal of Bilingualism, 15 (4): 426–445, 2011.

[20] B. Z. Pearson, S. C. Fernandez, V. Lewedeg, and D. Kimbrough. Oller, “The relation of input factors to lexical learning by bilingual infants,” Applied Psycholinguistics, 18 (1): 41–58, 1997.

[21] D. K. Dickinson and P.O. Tabors, Beginning literacy with language. Young children learning at home and school. Baltimore Brookes Publishing, 2001.

[22] J. Oakhill and K. Cain, “Introduction to comprehension development” in Children's comprehension problems in oral and written text: a cognitive perspective, K. Cain & J. Oakhill, Guilford Press, New York, pp. 3-40, 2007.

[23] C.S. Hammer, F.R. Lawrence, and A.W. Miccio, “Bilingual children’s language abilities and early reading outcomes in Head Start and kindergarten.” Language, Speech, and Hearing Services in Schools, 2007, doi: https://doi.org/10.1044/0161-1461(2007/025)

[24] S. Verdon, S. McLeod, and A. Winsler, “Language maintenance and loss in a population study of young Australian children,” Early Childhood Research Quarterly, 29 (2): 168–181, 2014.

[25] A. Lucero, “The Development of Bilingual Narrative Retelling Among Spanish–English Dual Language Learners Over Two Years,” Language, Speech, and Hearing Services in Schools, 49 (3): 607–621, 2018.

[26] K. E. Squires, M. J. Lugo-Neris, E. D. Peña, L. M. Bedore, T. M. Bohman, and R. B. Gillam, “Story retelling by bilingual children with language impairments and typically developing controls,” International Journal of Language & Communication Disorders, 49 (1): 60–74, 2013.

[27] E. Hoff and L. Naigles, “How Children Use Input to Acquire a Lexicon,” Child Development, vol. 73, no. 2, pp. 418–433, 2002.

[28] S. Jarvis, “Capturing the diversity in lexical diversity,” Language Learning, 63(1): 87–106, 2013.

[29] D. Malvern, B. Richards, N. Chipere, and P. Durán, Lexical Diversity and Language Development. London: Palgrave Macmillan UK, 2004.

[30] P. Escudero, G. Pino Escobar, M. Hernandez Gallego, C. Diskin-Holdaway & J. Hajek, “The Little Multilingual Minds corpus: Educators’ and children’s speech in heritage languages”, Workshop on community language corpora in Australia, Australian National University, Canberra, 2023.

[31] P. Escudero, G. Pino Escobar, C. Diskin-Holdaway, and J. Hajek. “Enhancing Heritage and Additional Language Learning in the Preschool Years: Longitudinal Implementation of the Little Multilingual Minds Program.” OSF Preprints. September 16, 20204. doi:10.31219/osf.io/rvjcg.

[32] P. Escudero, G. Pino Escobar, C. G. Casey & K. Sommer, “Four-year-old's online versus face-to-face word learning via eBooks”, *Frontiers in Psychology*, *12*, 450, 2021.

[33] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning (pp. 28492-28518). PMLR, 2013.

[34] Jasp: JASP Team (Version 0.18.3) [Computer software] 2024.

[35] J. K. Kruschke, “Bayesian estimation supersedes the t test.,” Journal of Experimental Psychology: General, 142(2): 573–603, 2013.

[36] Q. F. Gronau, A. Ly, and E.-J. Wagenmakers, “Informed Bayesian t-Tests,” The American Statistician, 74(2): 137–143, 2019.

Stop Oppositions in Warumungu: A Distributional and Acoustic Analysis

Mitchell Browne[1,2], Michael Proctor[1], Jane Simpson[3], Mark Harvey[4], Robert Mailhammer[5], and Harriet Carpenter[1]

[1]Macquarie University, [2]The University of Western Australia, [3]The Australian National University, [4]The University of Newcastle, [5]Western Sydney University

mitchell.browne@mq.edu.au

Abstract

Warumungu is a language of Central Australia with a reported contrast for stops. There has been no formal distributional or acoustic analysis of this contrast; this study begins to address this gap. Analysis of 439 stops produced by three speakers reveals that the consistent phonetic correlates of the manner contrast in morpheme-medial positions are duration and voicing. The primary distinction in stop manner is duration; voicing distinctions are predictable from duration.

Index Terms: stops, plosives, Warumungu, acoustic phonetics, Australian languages, duration, voicing

1. Introduction

The Warumungu (Pama-Nyungan) consonant inventory departs from the prototypical inventory of Australian languages (see [1], [2, pp. 605–615]) in having two intervocalically contrastive stop series at each place of articulation, illustrated in (1) and (2). The contrast has previously been described [3] as a distinction between short voiced (‘Mode 1’), and long voiceless stops (‘Mode 2’); however, the phonetic basis of the contrast has not been systematically examined.

	Mode 1	Mode 2
1)	/kantu/ ‘inside’	/nant:u/ ‘windbreak’
2)	/wacala/ ‘subsection name’	/cac:ula/ ‘dew, fog’

Warumungu stops have a long history of analysis. Previous linguistic analysis has recognised a phonetic manner opposition in stops. The first analysis was presented in 1953 by Arthur Capell, who observed stronger stop devoicing than in most Australian languages, as well as “a tendency to gemination of plosives”, but claimed that this was neither consistent nor phonemic [4]. Subsequent analyses have treated the distinction as phonemic: Ken Hale described a contrast between ‘tense’ and ‘lax’ stops [5], and Jeffrey Heath characterized the contrast as one of fortis/lenis [6].

One reason that analyses of this contrast differ is because Mode 2 stops occur in a restricted environment. Hale noted a general lengthening process of post-tonic consonants, which is only distinctive for stops (primary stress in Warumungu is generally word-initial; with secondary stress on non-word-final alternating syllables, with some lexical exceptions). This analysis was continued by Prithindra Chakravarti, who further observed a complementary distribution between long vowels and long consonants in stressed syllables, which he labelled “bimorism or bisyllabism” [7].

In addition to stress-related effects, there are also morphophonological interactions involving stops in Warumungu. Firstly, medial Mode 2 stops in disyllables become short when certain suffixes are added [3], [6]. For example, the Mode 2 stop in /ŋap:a/ is shortened when the dative suffix attaches: /ŋapa-ka/. Secondly, initial stops of these alternation-triggering suffixes themselves may be voiced or voiceless, depending on whether they were suffixed to disyllabic words or polysyllabic words, and on whether the suffix was monosyllabic or polysyllabic [5], [6]. In these environments, the ‘short’ stops are described as alternating between voiced (for disyllables) and voiceless (elsewhere) [3].

In sum, over the past seventy years, the existence of at least two series of oral stops has been recognised in Warumungu. What distinguishes the two series has been variously described as lax/lenis/short versus tense/fortis/long; however, no systematic quantitative phonetic analysis has been conducted to test these analyses and inform the characterization of this contrast. It is not clear what the primary phonological distinction between stops is; whether they are specified for more than one parameter; and how categorically the two stop series are distinguished.

1.1. Stop oppositions in Australian languages

Stop contrasts in Australian languages are rare [8]. Warlmanpa, a language neighbouring Warumungu, is reported to have a stop contrast based on length [9], however this requires further phonetic analysis. While it is not reported to be contrastive, medial stops in Warlpiri are reported to have high constriction duration in post-tonic position, with durations of approximately 100-140ms [10], [11]. In the Top End (the northern part of the Northern Territory), stop oppositions similar to the Warumungu system have been described in Bininj Gunwok [12], Burarra [13], Jawoyn [14], Murrinh-Patha [13], and Ngalakgan [15].

Cross-linguistically, stop contrasts characterized as voicing oppositions are typically realized through systematic differences in voice onset time (VOT) [16], [17]. In Murrinh-Patha and Ngan’gityemerri – languages of the Daly River region – VOT is the primary correlate of the stop manner opposition [18], [19]. In other languages of the Top End, VOT is not a consistent correlate of stop mode [12], [13], [14], [15], [20], so it does not appear that the stop manner opposition in these languages can be categorized as a voicing opposition. Analyses of these languages have proposed that the basis of the opposition is duration. In Bininj Gunwok for example, the mean length of lenis stops is 76 ms, compared to 161 ms for fortis stops [12, p. 173]. In other Top End languages, duration is not an independent predictor of stop type: Evans & Merlan [14, p. 216] propose that the phonological opposition in Jawoyn stops is determined by comparative quantitative weighting of both duration and voicing. No phonetic analysis of stops has yet been conducted to examine the basis of similar manner contrasts in Warumungu.

1.2. Research questions

The aim of this study is to examine the distributional and acoustic phonetic properties of Warumungu stops, to determine:

1. the relative lexical distribution of Mode 1 and Mode 2 stops inside monomorphemic words
2. whether duration differs between Mode 1 and Mode 2 stops
3. whether voicing differs between Mode 1 and Mode 2 stops
4. how duration and voicing interact

Based on the answers to these questions, we will address the following hypotheses about the Warumungu stop contrast: (H1) duration alone is phonologically specified; (H2) voicing alone is phonologically specified; and (H3) both duration and voicing are independently specified in the phonology.

2. Method

2.1. Distributional analysis

To establish the distribution of stops of different type, we examined an existing word list of Warumungu. Because inflection influences stop mode, we excluded any morphologically complex entries (e.g. verbs, which require an inflection to be well-formed) and words which are orthographically recognised as complex (e.g. complex verb constructions), as well as bound morphemes. After these exclusions, the total number of words considered in our distributional analysis was 804.

2.2. Acoustic analysis

Data used for acoustic analysis is drawn from two sources, both comprising careful speech. The first source is an audio archive containing acoustic recordings of a single adult female speaker (age 40+) producing citation form examples for the Warumungu Picture Dictionary [22], either in isolation or short sentences. The second data source for the phonetic study consists of acoustic recordings made during experimental fieldwork conducted around Tennant Creek in 2023 and 2024. These data were elicited from two adult female (ages 50+ and 60+), both native speakers of Warumungu. Participants were shown a series of 32 images corresponding to Warumungu words containing target segments, checked by speakers as to whether they represented the intended target word.

Participants were instructed to say the corresponding Warumungu word twice for each image. Each participant was shown the series of images in a different (random) order. Each image was displayed until the participant provided a response. Responses were recorded using a Zoom H5 with an LMF-1 Lavalier microphone. Speech audio was recorded at a sampling rate of 44.1 kHz and saved as 16-bit uncompressed WAV files. Participants were remunerated for their time.

970 Warumungu words containing stops were identified in the data sources, of which 461 met the criteria for acoustic analysis: only medial stops in tautomorphemic words of at least two syllables were analyzed.

Matlab [23] tools were developed for data inspection, audition, and analysis, to facilitate systematic processing and acoustic segmentation of the dataset. Medial stop intervals were located in each word using a semi-automatic algorithm. The interval of occlusion was selected on a waveform/spectrogram plot by a phonetically trained analyst (Author 1) and a supervised student (Author 6). Occlusion boundaries were estimated automatically from amplitude and energy thresholds calculated over overlapping 20 ms windows, and manually corrected when necessary. Stop duration was calculated over the total interval of occlusion including any carryover voicing and negative voice onset intervals (Figure 1). Voicing was estimated from harmonic ratios (HR), which calculate the distribution of energy across frequency bands [24], [25]. Voiced speech is characterized by greater concentration of energy at multiples of the fundamental frequency (HR → 1), while voiceless/devoiced speech is produced with more even spectral distribution of energy, resulting in lower harmonic ratios (HR → 0). Harmonic ratios were calculated over 25 ms overlapping windows throughout each word, and the degree of voicing for each stop was calculated as the mean HR over the occlusion interval. Using mean HR allows for consistent analysis between stops, particularly where clear landmarks are not identifiable (e.g. approximant realisations).

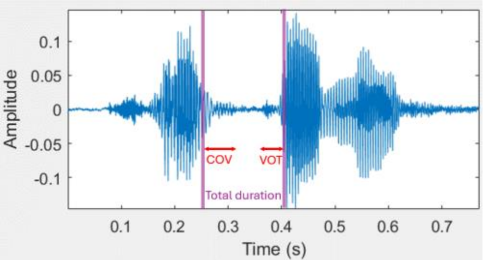


Figure 1: *Total duration example of* /cap:ina/ *'bearded dragon'.*

Each token in the acoustic analysis was coded for place of articulation; mode (based on the current practical orthography used in Warumungu dictionaries in which stops are written with one letter (Mode 1) or two letters (Mode 2) , e.g. *kantu* 'inside', *nanttu* 'windbreak'); speaker; word; total syllable count of word; preceding morae (N); mean HR; and duration. Tokens with excessive background noise were excluded. Final token counts are shown in Table 1.

Table 1: *Number of stops included in acoustic analysis.*

Place of articulation	Mode 1	Mode 2
Bilabial	77	33
Alveolar	32	25
Retroflex	44	26
Palatal	63	36
Velar	79	46
Total	295	166

2.3. Statistical analysis

Outliers in which mean HR deviated more than two standard deviations from the mean for each stop mode were excluded from statistical analysis. 22 of the original 461 tokens were excluded by this criterion, leaving 439 stops for analysis.

We examined stop duration, voicing, and the interaction between duration and voicing using linear mixed-effects models in R [26], [27], [28]. Stop duration (3) was modelled as a function of stop mode with interactions for place of articulation, and random effects of word (intercepts & slopes) and speaker (intercepts only). Stop voicing (4) was modelled as a function of stop mode (baseline Mode 1) with interactions for place of articulation (baseline Alveolar), and random effects of word (intercepts only) and speaker (intercepts & slopes). The most descriptive convergent model was chosen for each variable.

3) duration ~ Mode*PoA + (Mode|Word) + (1|Speaker)

Table of Contents for this manuscript

4) voicing ~ Mode*PoA + (1|Word) + (Mode|Speaker)

To examine the relationship between duration and voicing, we modelled voicing (mean HR) as a dependent variable and duration as the independent variable. We also examined the overlap in the distribution of tokens (with respect to duration and voicing), calculated with the R package OVERLAPPING [29]. Plots were generated with the PLOTLY package [30] in R.

3. Results

3.1. Distributional analysis

MODE 1 stops are relatively unrestricted, able to occur in the onset of any syllable. MODE 2 stops are primarily found in disyllables following primary stress.

Considering only words which have a stop in post-tonic position (n=407), 246 (60%) are documented as MODE 1, and 161 (40%) as MODE 2 (Figure 2). When further restricting this to disyllabic words (n=145) the ratio is reversed: 50 (34%) disyllabic words have a MODE 1 stop following primary stress; and 95 (66%) words have a MODE 2 stop in this environment. At each place of articulation, both stop types occur intervocalically, and as the second member of a cluster, i.e. V.**C**V and VC.**C**V respectively. MODE 1 stops can rarely occur as the first member of a cluster, whereas MODE 2 stops are not evident in this environment.

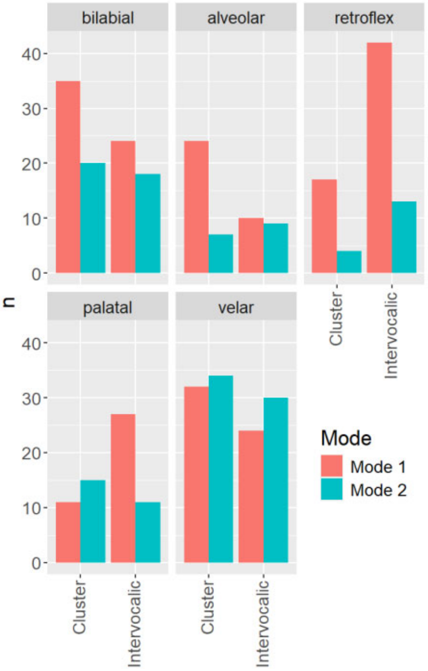


Figure 2: *Distribution of stops found in second syllable onsets, grouped by stop mode and environment.*

The distribution and location of MODE 2 stops in words of different length are summarized in Table 2. 89% of MODE 2 stops are found following a primary stress (i.e. 1 preceding syllable), the majority of which are found in disyllable words; and 7% following secondary stress (i.e. 3 preceding syllables). The remaining 4% of MODE 2 stops are found in the onset of the third syllable, e.g. /kuɭaɲ**c**:ari/ 'diamond dove'. In addition to stress pattern restrictions, MODE 2 stops cannot follow long vowels; and cannot be the first member of a consonant cluster (though they may occur in a consonant cluster following a nasal or a continuant). There are no (morphologically simple) words which contain multiple MODE 2 stops.

It is also notable that Warumungu is distinct from most Australian languages with a stop opposition in that the stop contrast in Warumungu is active following nasals [8], in addition to following continuants.

Table 2: *Distribution of Mode 2 stops by word length and location within word.*

PRECEDING	TOTAL SYLLABLES IN WORD					
SYLLABLES	2σ	3σ	4σ	5σ	6σ	7σ
1σ	95	48	14	1	1	2
2σ			4			
3σ			12	1		

3.2. Duration

Duration of MODE 2 stops is significantly longer than that of MODE 1 stops across all places of articulation (β = 136, t(424) = 11.6, p < .001). Bilabial and palatal PoAs were significant. Overall mean duration of MODE 1 stops is 43 ms (s.d. 23), compared to 170 ms (s.d. 41) for MODE 2 stops – a duration ratio of approximately 1:4. Figure 3 shows mean durations for each stop mode across each place of articulation. Similar duration differences were found between MODE 1 and MODE 2 stops occurring in post-tonic position in di- or tri-syllabic roots (β = 137.9, t(221) = 10.98, p < .001).

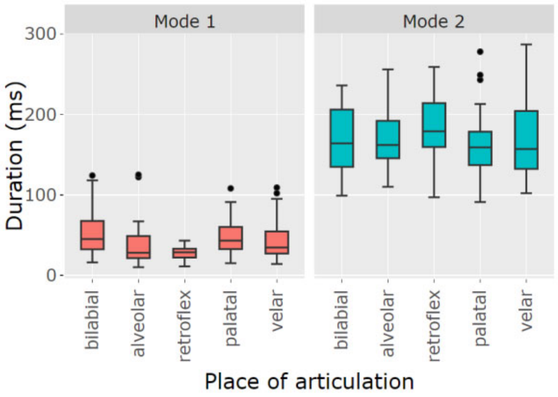


Figure 3: *Stop duration by mode and place of articulation.*

3.3. Voicing

MODE 2 stops have significantly lower mean HR compared to MODE 1 stops across all places of articulation (β = –0.19, t(424) = –3.5, p < .001). There was a significant interaction between palatal place of articulation and stop mode (β = –0.04, t(424) = –2.2, p = .003). The overall mean HR for MODE 1 stops is 0.66 (s.d. 0.07), and 0.47 (s.d. 0.10) for MODE 2 stops. Higher harmonic ratios are characteristic of voiced speech, so this indicates that MODE 1 stops are more voiced than MODE 2 stops. Mean HR for each stop mode across each place of articulation is shown in Figure 4. Similar differences in voicing were found between MODE 1 and MODE 2 stops occurring in post-tonic position in di- or tri-syllabic roots (β = –0.22, t(221) = –4.0, p < .001).

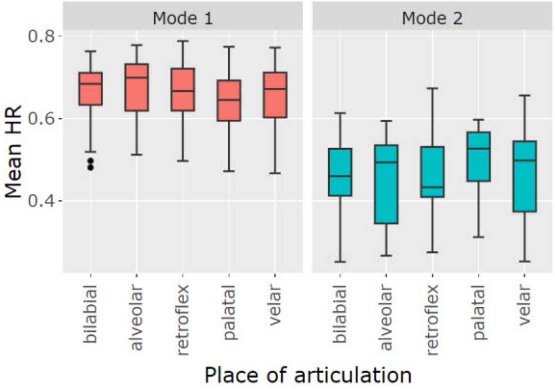


Figure 4: *Mean harmonic ratio for Mode 1 (left) and Mode 2 (right) stops, by place of articulation.*

3.4. Interaction between duration & voicing

Duration was a significant factor predicting voicing (β = –0.001, t(433) = –15.7, p < .001). The Pearson correlation between duration and voicing is –0.73, independent of stop mode (Figure 5). Density plots (Figure 6) reveal bi-modal distributions for both duration and mean HR (grouped by stop mode). The stop modes have an overlapping area of 0.06 for duration and 0.31 for mean HR.

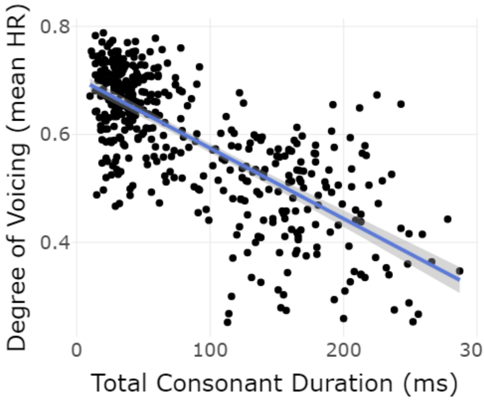


Figure 5: *Total stop duration as a predictor of degree of voicing.*

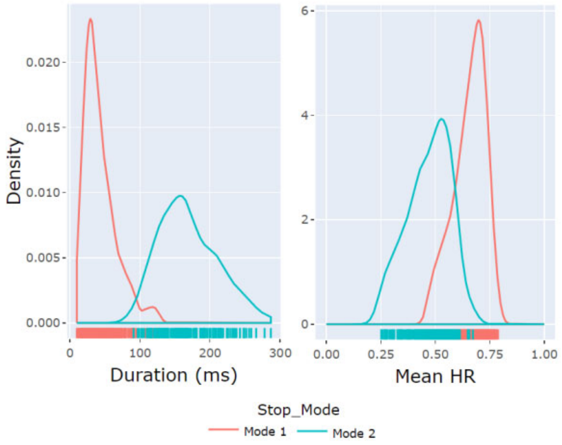


Figure 6: *Distribution of duration (left) and mean HR (right) across Mode 1(red) and Mode 2 (aqua) stops.*

4. Discussion

These data reveal that the contrast between stop modes in Warumungu is phonetically realised with duration and voicing. These phonetic contrasts are active across all places of articulation, for stops occurring intervocalically and in clusters.

Our research questions relate to whether one or both of these parameters is phonologically specified. Given the strong correlation between duration and voicing, we do not find support for H3, as duration and voicing reliably predict one another in root-medial environments. H1 is supported by the distributional and phonetic evidence. The opposition is not active word-initially, and consonant sequences are not found word-initially, and so the gemination analysis aligns with these phonotactic expectations. In contrast, a voicing opposition would be expected to occur in these environments.

Based on the density distributions, duration appears more distinctive, with very little overlap between the stop distributions (6%; in the 91-125ms duration range); compared with mean harmonic ratio (30%). This is quite unlike other analyses of stop duration, such as that of Jawoyn, in which duration was an important factor, yet there was considerable overlap between the duration of short and long tokens, and thus other factors are necessary to account for the stop distinction [14]. This is also evidenced by the higher stop duration ratio (1:4) compared to those reported for other languages, e.g. 1:2 in Bininj Gunwok [12], and (up to) 1:3 cross-linguistically [31]. Similarly, Warumungu Mode 2 stops are longer than (phonetically) long stops in Warlpiri [10], [32].

Taken together, the results provide strong evidence in favour of H1, in which duration is the sole phonological specification distinguishing Mode 1 stops and Mode 2 stops in Warumungu. In this analysis, stops have the same gestural specification, and differ in the number of linked timing units.

This analysis has a number of advantages over H2 (in which voicing is the sole phonological specification). Primarily, the overlap between modes is minimal for duration, compared to voicing. It avoids the proliferation of contrastive features in the Warumungu inventory: [voice] is not phonologically active in Warumungu, and the phonetic realisation of voicing arises through constriction duration. This is supported by the strong correlation between duration and mean harmonic ratio, which suggests that voicing is predictable from the implementation of duration. However, many acoustic properties of Warumungu stops remain unexplored. This includes the complex morphophonological interactions, as well as more fine-grained voicing parameters (e.g. VOT) and variation in the actual manner of articulation of (underlying) stops.

5. Conclusion

This study has found robust support for a stop opposition in Warumungu. Phonetically, this opposition is realised by duration and voicing. Phonologically, this opposition is reliably differentiated by total occlusion duration. The difference in voicing can be captured by the strong negative correlation between duration and voicing, rather than being phonologically specified. The data suggest a geminate analysis for Warumungu stops, rather than a fortis/lenis analysis. However, the geminate analysis requires an expansion of the syllable inventory of Warumungu—specifically either the existence of coda clusters (e.g. '*maŋk.ka* 'hole') or complex onsets (e.g. '*maŋ.kka*)—which warrants further consideration, as neither of these cluster types are otherwise evident in Warumungu.

6. Acknowledgements

We are grateful to the Warumungu speakers who teach us about their language. We also thank the two anonymous reviewers for their detailed comments on this paper. We gratefully acknowledge funding support from the Australian Research Council grant DP220102925 and from the 2024 Macquarie University ECR Enabling Scheme.

7. References

[1] J. Fletcher and A. Butcher, "Sound patterns of Australian Languages," in *The Languages and Linguistics of Australia*, H. Koch and R. Nordlinger, Eds., Berlin, Boston: De Gruyter Mouton, 2014, pp. 91–138. doi: 10.1515/9783110279771.91.

[2] R. M. W. Dixon, *Australian languages : their nature and development*. New York: Cambridge University Press, 2002. [Online]. Available: http://www.loc.gov/catdir/toc/cam024/2001037958.html

[3] J. Simpson, "Warumungu (Australian – Pama-Nyungan)," in *The Handbook of Morphology*, A. M. Zwicky and A. Spencer, Eds., 2017, pp. 707–736. doi: 10.1002/9781405166348.ch32.

[4] A. Capell, "Notes on the Waramunga language, Central Australia," *Oceania*, vol. 23, no. 4, pp. 297–311, Jun. 1953, doi: 10.1002/j.1834-4461.1953.tb00199.x.

[5] K. Hale, *Warramunga notes. Sandy Nandy and George Bruce from Tennant Creek*. MS 863. Australian Institute of Aboriginal and Torres Strait Islander Studies, 1959.

[6] J. Heath, *Warramunga grammatical notes, based on fieldwork in April/June 1977*. PMS 3804. Canberra: Australian Institute of Aboriginal and Torres Strait Islander Studies, 1977.

[7] P. Chakravarti, *WaRumungu*. PMS 332. Canberra: Australian Institute of Aboriginal and Torres Strait Islander Studies.

[8] E. R. Round, "Segment inventories," in *The Oxford Guide to Australian Languages*, 1st ed., C. Bowern, Ed., Oxford University PressOxford, 2023, pp. 96–105. doi: 10.1093/oso/9780198824978.003.0010.

[9] M. Browne, *A Grammar of Warlmanpa*. Canberra: ANU Press, 2024. doi: 10.22459/GW.2024.

[10] C. Pentland and M. Laughren, "Distinguishing prosodic word and phonological word in Warlpiri: prosodic constituency in morphologically complex words," *Proceedings of the 2004 Conference of the Australian Linguistics Society*, no. 1995, p. 17, 2004.

[11] R. L. Bundgaard-Nielsen and C. O'Shannessy, "Voice onset time and constriction duration in Warlpiri stops (Australia)," *Phonetica*, vol. 78, no. 2, pp. 113–140, Apr. 2021, doi: 10.1515/phon-2021-2001.

[12] H. Stoakes, "An Acoustic and Aerodynamic Analysis of Consonant Articulation in Bininj Gun-wok," PhD, The University of Melbourne, Melbourne, 2013.

[13] A. Butcher, "'Fortis/lenis' revisited one more time: the aerodynamics of some oral stop contrasts in three continents," *Clinical Linguistics & Phonetics*, vol. 18, no. 6–8, pp. 547–557, Sep. 2004, doi: 10.1080/02699200410001703565.

[14] B. Evans and F. Merlan, "Stop contrasts in languages of Arnhem Land: From the perspective of Jawoyn, Southern Arnhem Land," *Australian Journal of Linguistics*, vol. 24, no. 2, pp. 185–224, Oct. 2004, doi: 10.1080/0726860042000271825.

[15] B. J. Baker, *Word structure in Ngalakgan*. in CSLI publications. Stanford, Calif: Center for the Study of Language and Information, 2008.

[16] C. Henton, P. Ladefoged, and I. Maddieson, "Stops in the World's Languages," *Phonetica*, vol. 49, no. 2, pp. 65–101, Mar. 1992, doi: 10.1159/000261905.

[17] L. Lisker and A. S. Abramson, "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements," *WORD*, vol. 20, no. 3, pp. 384–422, Jan. 1964, doi: 10.1080/00437956.1964.11659830.

[18] N. Reid, *Ngan'gityemerri: a language of the Daly River region, Northern Territory of Australia (Outstanding Grammars from Australia 6)*. Munich: Lincom Europa, 2011.

[19] J. Mansfield, *Murrinhpatha morphology and phonology (Pacific Linguistics 653)*. Berlin & Boston: De Gruyter Mouton, 2019.

[20] R. Mailhammer, S. Sherwood, and H. Stoakes, "The inconspicuous substratum: Indigenous Australian languages and the phonetics of stop contrasts in English on Croker Island," *English World-Wide*, vol. 41, no. 2, pp. 162–192, 2020.

[21] W. R. Leben, "A Metrical Analysis of Length," *Linguistic Inquiry*, vol. 11, no. 3, pp. 497–509, 1980.

[22] S. Disbray and D. N. Stokes, *Warumungu Picture Dictionary*. Alice Springs, N.T.: IAD Press, 2005.

[23] T. M. Inc, *MATLAB version: 24.1.0 (R2024a)*. (2024). The MathWorks Inc., Natick, Massachusetts, United States. [Online]. Available: https://www.mathworks.com

[24] X. Sun, "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio," *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, p. I-333-I–336, 2002.

[25] T. Drugman and A. Alwan, "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics," 2020, *arXiv*. doi: 10.48550/ARXIV.2001.00459.

[26] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2024. [Online]. Available: https://www.R-project.org/

[27] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[28] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017, doi: 10.18637/jss.v082.i13.

[29] M. Pastore, P. A. D. Loro, M. Mingione, and A. Calcagni', *overlapping: Estimation of Overlapping in Empirical Distributions*. 2022. [Online]. Available: https://CRAN.R-project.org/package=overlapping

[30] C. Sievert, *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020. [Online]. Available: https://plotly-r.com

[31] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*. Oxford: Blackwell Publishers, 1996.

[32] R. L. Bundgaard-Nielsen and C. O'Shannessy, "Voice Onset Time and Constriction Duration in Warlpiri Stops (Australia)," in *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, 2019.

Phonetic Evidence for Fortis-lenis Contrast in Kufo Pulmonic Plosives

Shubo Li, Rosey Billington

Australian National University

shubo.li@anu.edu.au, rosey.billington@anu.edu.au

Abstract

The Kufo language of Sudan has pulmonic and non-pulmonic consonants. Within the pulmonic consonants, it has until recently been unclear how many plosive series are contrastive. Early studies suggest that there is a voicing contrast, and perhaps a length contrast as well, but more recent work proposes the possibility of a fortis-lenis type contrast which involves both voicing and length. This paper examines the phonetic evidence for the proposed contrast, based on data collected with one speaker. Results for Voice Onset Time and closure duration support the analysis of two pulmonic plosive series in Kufo, and align with an interpretation of fortis vs. lenis as the basis of the contrast.

Index Terms: Kufo, plosives, fortis-lenis, manner of articulation, closure duration, Voice Onset Time

1. Introduction

Kufo[1] is a variety of the Kanga language, which is traditionally spoken in the Nuba Mountains in South Kordofan, Sudan. Kanga is classified as part of the Kadugli-Krongo language family, and with approximately 8,000 speakers, it is viewed as severely endangered [1][2]. All of the Kadugli-Krongo languages are understudied, and for Kufo, previous work is largely limited to phonological observations in the context of comparative discussions. The current study is part of a wider documentation project involving the only diasporic Kufo speaker residing in Australia, and examines the phonetic evidence for a proposed fortis-lenis contrast in Kufo pulmonic plosives.

1.1. Kufo and Kadugli-Krongo pulmonic plosives

Early work on Kufo and closely related varieties points to plosive contrasts at five supralaryngeal places of articulation, including non-pulmonic plosives /ɓ, ɗ/, and at least one pulmonic plosive series. In [3][4], contrastive voiced and voiceless pulmonic plosives are proposed, and 'most consonants' reportedly also occur as geminates. Other work instead proposes a single pulmonic plosive series, with no voicing contrast, and phonetic voicing in intervocalic contexts [5][6]. [5] also notes that all plosives can occur as geminates, but does not analyse these as contrastive, and they are not mentioned in [6]. Across the Kadu language family, it is also unclear what the typological and historical patterns of plosives contrasts are [5].

More recent work proposes that the Kufo pulmonic plosives exhibit a 'fortis' vs. 'lenis' type contrast [7]. The proposed Kufo plosive inventory (Table 1) has supralaryngeal pulmonic plosives including four lenis-fortis pairs at bilabial /p, pː/, dental /t, tː/, retroflex /ʈ, ʈː/, and velar /k, kː/ places of articulation, plus the lenis palatal plosive /c/ which does not have a fortis counterpart. This proposal is based on phonological evidence that while short voiced and voiceless plosive phones both occur, they do not occur in the same environments, and correspond to a single plosive series. Based on auditory impressions, these 'lenis' plosives are realised as voiceless in word-initial position and voiced in word-medial position. The allophonic variation is evident in the process of pluralisation [7]. For example, the lexical item 'stick' is /tɔlɔ/ [tɔlɔ] in its singular form, where the lenis plosive is phonetically voiceless in initial position, but with the addition of the plural prefix /na-/, the stem-initial plosive is word-medial and phonetically voiced, as in /natɔlɔ/ [nadɔlɔ]. In comparison, the fortis plosives only occur in word-medial position, and are always voiceless, and impressionistically longer than the lenis plosives. There is clear evidence for contrast between the fortis and lenis plosives, e.g., /mutu/ [mudu] 'wine waste' and /mutːu/ [mutːu] 'horse'. However, the phonetic cues to the proposed fortis-lenis contrast among Kufo pulmonic plosives have not yet been examined phonetically.

Table 1: *Proposed Kufo plosive inventory [7].*

	bil.	den.	ret.	pal.	vel.	glo.
lenis	/p/	/t/	/ʈ/	/c/	/k/	/ʔ/
fortis	/pː/	/tː/	/ʈː/		/kː/	
implosive	/ɓ/	/ɗ/				

1.2. Fortis-lenis contrasts in other languages

The terms 'fortis-lenis' were adopted in [7] because the perceived nature of the contrast in pulmonic plosives aligns with contrasts described similarly in other languages (e.g.[10]), noting, however, that the difference between 'fortis-lenis' (also 'tense-lax') vs. 'geminate-singleton' is not clear cut. Fortis-lenis contrasts are generally described as relating to consonantal strength, involving differences in respiratory and articulatory energy and with a range of language-specific acoustic correlates [8]. For example, in Korean, fortis, lenis, and aspirated plosives are distinguished by acoustic and aerodynamic parameters including Voice Onset Time (VOT), fundamental frequency (f0), intraoral air pressure and air flow [9]. In varieties of Germanic languages such as English and German, a primary correlate of fortis-lenis contrasts is VOT, with long-lag VOT for fortis plosives and short-lag VOT for lenis. Reported secondary correlates include closure duration, f0, burst intensity and often voicing intervocalically (e.g. [24]).

In other languages, the primary acoustic correlate of contrasts described as fortis-lenis is duration, which, being the primary correlate for geminate-singleton contrasts, presents challenges in determining the most appropriate phonological descriptors [11] [22]. In Swiss German, fortis plosives have longer closure durations than lenis plosives, and the two series do not differ in VOT [21]. In Bininj Gun-wok, one of many Australian languages described as having a fortis-lenis contrast, fortis plosives are around twice as long as lenis plosives, and also have

[1] ISO 639-3: kcp, Glottolog: kang1288

higher intro-oral pressure, distinguishing them from morphological geminates [12]. Both lenis and fortis plosives in Bininj Gun-wok have short-lag VOT, and word-medial lenis plosives are often realised as fricatives or approximants. A number of Oto-Manguean languages are also described as having fortis-lenis contrasts drawing on duration. In Itunyoso Trique, fortis obstruents have longer closures than lenis obstruents, and also exhibit preaspiration [13]. Lenis obstruents also show variable voicing and spirantization, which [13] argues can be explained by their reduced durations, suggesting that the consonant contrast may be best considered one of length rather than strength. Arguments in the opposite direction can also be found; for example, the length contrast attested for Somali voiced plosives is primarily realised as a manner contrast, with short voiced plosives largely produced as approximants [23]. It is clear that understanding the phonetic and phonological typology of strength and length contrasts requires more detailed studies of diverse languages.

2. Research aim

This study aims to examine the phonetic evidence for the proposed fortis-lenis contrast in Kufo pulmonic plosives. The following questions will be addressed: What are the phonetic realisations of the two pulmonic plosive series in Kufo, in different word positions? Do the acoustic cues to the contrasts support an interpretation of fortis vs. lenis as the basis of the contrast?

3. Method

3.1. Participant

The speech data for this study was collected with Haroun Kafi, the only Kufo speaker residing in Australia. Haroun was born in the 1960s, grew up in Sudan, and currently resides in rural Victoria. Besides Kufo, Haroun also speaks Sudanese Arabic and English.

3.2. Materials and procedures

A wordlist of 54 disyllabic words was developed based on lexical data and phonological analyses in recent work [7]. Table 2 presents some of the lexical items included in the wordlist, based on phoneme and word position. The words in the wordlist predominantly have a CV.CV structure and short vowels only, to the extent which current data allows.

Table 2: *Example words included in the wordlist.*

lenis	INI	translation	MED	translation
/p/	/paʔja/	‘all’	-	-
/t/	/tafa/	‘have’	/mutu/	‘wine waste’
/ʈ/	/ʈiko/	‘dam’	/taʈe/	‘cut’
/c/	/coːno/	‘dig’	/teca/	‘wake’
/k/	/kaʈɛ/	‘wings’	/kika/	‘where’
fortis	INI	translation	MED	translation
/pː/	-	-	/napːa/	‘fathers’
/tː/	-	-	/mutːu/	‘horse’
/ʈː/	-	-	/ʈaʈːo/	‘woodpecker’
/kː/	-	-	/tukːu/	‘write’

Data collection was conducted in a quiet room at the speaker’s home. Audio was recorded with a Zoom H6 audio recorder and a Røde NT3 cardioid microphone, at an archival sampling rate of 96kHz and 24-bit depth. Lexical items in the wordlist were elicited with English verbal prompts in a random order. Each word was produced 5 times consecutively within the utterance-medial frame *aʔa nɪkːi* ... *ɓɪtɛnɪ* ‘I say ... today’.

3.3. Data processing and analysis

The sound files were downsampled to 44.1kHz and 16-bit depth for acoustic analysis, and segmented and annotated in Praat [14]. VOT and closure duration are the primary measures of interest in this study, depending on segment position and phonetic realisation. For word-initial and word-medial plosives, VOT was segmented based on the onset of the release burst and the onset of periodicity for the following vowel (see Figure 1). Closure duration for plosives in word-medial position was segmented based on the last glottal pulse of the preceding vowel and the onset of the release burst of the target plosive. For plosives in word-initial position, closure duration was not annotated, as the speaker typically produced a short pause before the target word and the onset of the closure for the phonetically voiceless plosives in this word position could not be reliably segmented. Intervocalic plosives were sometimes phonetically realised as approximants. As such, VOT and closure duration are not reported for these cases.

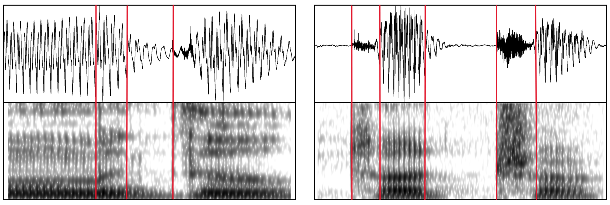


Figure 1: */mutu/ ‘wine waste’ & /ʈoʈː/ ‘woodpecker’.*

Based on the .wav files and paired .TextGrids, a hierarchical speech database was created using the EMU Speech Database Management System [15]. In total, the database consists of 377 consonant tokens, including 219 phonemic lenis plosives in word-initial position, 88 phonemic lenis plosives in word-medial position, and 70 phonemic fortis plosives in word-medial position. A summary of the number of tokens in this dataset is presented below in Table 3. For the relevant consonants realised phonetically as plosives, measures of VOT and closure duration were extracted and analysed with R [16], using the emuR package [17]. The lenis bilabial plosive /p/ has an extremely low functional load and occurs in one lexical item in initial position only based on available data.

Table 3: *Number of tokens, by phoneme and word position.*

lenis	INI	MED	total	fortis	MED
/p/	5	-	5	/pː/	10
/t/	117	21	138	/tː/	15
/ʈ/	26	20	46	/ʈː/	15
/c/	15	16	31		
/k/	56	31	87	/kː/	30
total	219	88	307		70

4. Results

4.1. Phonetic realisation

A summary of the realisations of all consonant tokens in this dataset is given in Table 4. Phonetic catgorisations are based on auditory impressions with reference to waveforms and corresponding spectrograms. Lenis plosives in word-initial position are mostly realised as phonetically voiceless plosives (95% of the time), though 47% of tokens of the lenis palatal plosive /c/ and 5% of tokens of the lenis velar plosive /k/ are realised as approximants. Lenis plosives in word-medial position are always phonetically voiced, and realised more often as approximants [ɹ, j, ɰ] (63%) than plosives [d, ɟ, g] (37%). Phonemic plosives are more likely to be realised as approximants as the place of articulation goes back, except the retroflex plosive /ʈ/, which is predominantly realised as the voiced plosive [ɖ] (95%). Fortis plosives in word-medial position are always realised as phonetically voiceless plosives, and never as approximants.

Table 4: *Phonetic realisation of lenis and fortis plosives in word-initial and -medial position.*

lenis	INI		MED		fortis	MED	
	plo.	appr.	plo.	appr.		plo.	appr.
/p/	5	0	-	-	/pː/	10	0
/t/	117	0	9	12	/tː/	15	0
/ʈ/	26	0	19	1	/ʈː/	15	0
/c/	8	7	4	12			
/k/	52	3	1	30	/kː/	30	0
total	208	10	33	55		70	0

The distributions in Table 4 show that the lenis plosives are more variable in terms of voicing and manner of articulation, with more (voiced) approximant than (voiceless) plosive phonetic realisations word-medially, while the fortis plosives, which only occur word-medially, do not show the same variation. The next section presents durational measures for the plosive phonemes that are phonetically realised as plosives.

4.2. Durational measures

4.2.1. Closure duration

Closure duration measures for lenis and fortis plosive phonemes realised as phonetic plosives in word-medial position are shown in Table 5 and Figure 2. On average, fortis plosives have a closure duration of 117ms, approximately 1.9 times the average duration of lenis plosives, 62ms. Based on a linear mixed-effects model with closure duration as the dependent variable, fortis/lenis as the independent variable, and word and place of articulation as random effects, the difference in closure duration between fortis and lenis phonetic plosives in word medial position is statistically significant (p<0.001***).

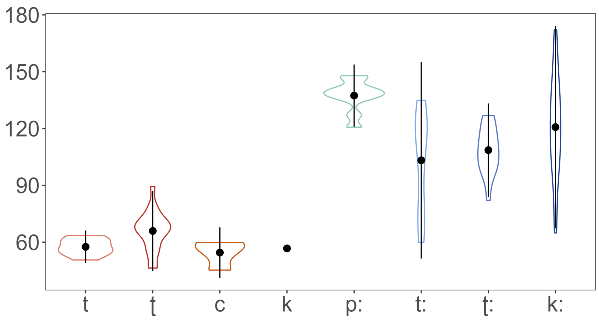


Figure 2: *Closure duration (ms) of lenis and fortis plosive phonemes in word-medial position, when phonetically realised as plosives.*

Table 5: *Closure duration (ms) of lenis and fortis plosive phonemes in word-medial position, when phonetically realised as plosives.*

lenis	phone	mean	sd	fortis	phone	mean	sd
				/pː/	[pː]	137	8
/t/	[d]	58	4	/tː/	[tː]	103	26
/ʈ/	[ɖ]	66	11	/ʈː/	[ʈː]	109	12
/c/	[ɟ]	54	7				
/k/	[g]	57	NA	/kː/	[kː]	121	27
total		62	10			117	24

4.2.2. Voice Onset Time

As seen in Table 4, when lenis plosives in word-initial position are not realised as approximants, they are realised as voiceless plosives [p, t, ʈ, c, k], with positive VOT as shown in Table 6 and Figure 3. The average VOT across different places of articulation is 40ms, which aligns with tendencies for languages in which plosives are described as unaspirated or weakly aspirated, with short-lag positive VOT (e.g. [19], [20]). For lenis plosive phonemes in word-initial position that are phonetically realised as plosives, there is a tendency for VOT to get longer as the place of articulation goes back, apart from for the retroflex, which has the longest VOT, but VOT differences by place are not statistically significant according to a linear mixed-effects model with VOT as the dependent variable, place as the independent variable, and word as a random effect.

Table 6: *VOT (ms) of lenis plosive phonemes in word-initial position, when phonetically realised as plosives.*

lenis	phone	mean	sd
/p/	[p]	30	8
/t/	[t]	37	10
/ʈ/	[ʈ]	52	26
/c/	[c]	34	31
/k/	[k]	46	14
total		40	16

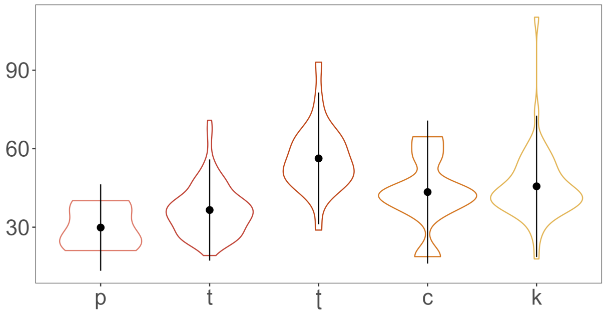


Figure 3: *VOT (ms) of lenis plosive phonemes in word-initial position, when phonetically realised as plosives.*

Table 4 showed that in word-medial position, lenis plosive phonemes are always voiced when phonetically realised as plosives, and this is evidenced by negative VOT values in Table 7 and Figure 4. (Note that only one /k/ token in word-medial position is realised as a plosive [g].) Given that they are fully voiced, the VOT of medial lenis plosives is the same as their closure duration. The average negative VOT of -62ms is different to the positive VOT for lenis plosives in word-initial position. For fortis plosive phonemes, which are always realised as phonetically voiceless plosives [pː, tː, ʈː, kː], positive VOT values are shown in Table 7 and Figure 4. The average VOT across different places of articulation is 50ms, slightly longer than for word-inital lenis plosives. VOT for fortis plosives in word-medial position shows a similar pattern to the lenis plosives in word-initial position, with a tendency for VOT to get longer as the place of articulation goes back, apart from for the retroflex, which has the longest VOT. Based on a linear mixed-effects model with VOT as the dependent variable, fortis/lenis as the independent variable, and word and place of articulation as random effects, the difference in VOT for fortis compared to lenis phonemes realised as plosives in word medial position is statistically significant (p<0.001***).

Table 7: *VOT (ms) of lenis and fortis plosive phonemes in word-medial position, when phonetically realised as plosives.*

lenis	phone	mean	sd	fortis	phone	mean	sd
				/pː/	[pː]	29	7
/t/	[d]	-58	4	/tː/	[tː]	44	11
/ʈ/	[ɖ]	-66	11	/ʈː/	[ʈː]	71	5
/c/	[ɟ]	-54	7				
/k/	[g]	-57	NA	/kː/	[kː]	51	10
total		-62	10			50	16

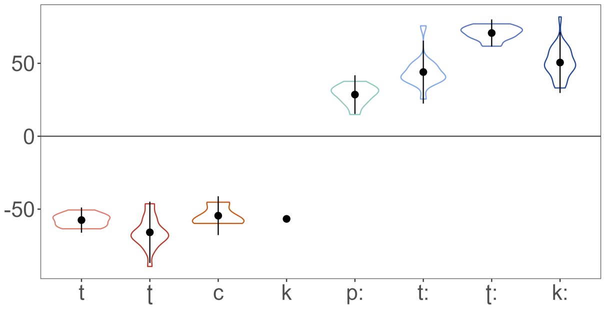


Figure 4: *VOT (ms) of lenis and fortis plosive phonemes in word-medial position, when phonetically realised as plosives.*

5. Discussion and conclusion

This study addressed two research questions. The first related to the phonetic realisation of Kufo pulmonic plosives, following recent proposals that Kufo has two contrastive pulmonic plosive series viewed as 'fortis' and 'lenis' [7]. Phonetic results provide supporting evidence for the phonological analysis of two plosive series, rather than three (voiced, voiceless, voiceless geminate [3][4]) or one (voiceless [6]), with different phonetic realisations depending on phonological environment. The differences in phonetic realisations involve voicing, duration, and manner of articulation. Fortis plosive phonemes, which only occur word-medially, are always realised as voiceless plosives, and never as approximants. Lenis plosive phonemes in word-initial position are mostly realised as voiceless plosives and occasionally approximants. Lenis plosive phonemes in word-medial position are always voiced, and are more likely to be realised as approximants than plosives, except the retroflex /ʈ/, which is almost always phonetically realised as a plosive. Compared to the fortis plosive phonemes, the lenis plosive phonemes in Kufo are more variable regarding their phonetic realisations.

The acoustic correlates investigated in this study include closure duration and VOT. In word-medial position, closure duration for fortis plosives is significantly longer than for lenis plosives, with fortis plosives on average 1.9 times longer. For VOT, lenis plosives in word-initial position have positive (short-lag) VOT, whereas lenis plosives in word-medial position have negative VOT. The fortis plosives, which occur in word-medial position only, have positive VOT (also short-lag). The VOT difference between fortis and lenis plosives in word-medial position is significantly different. For both lenis plosives in word-initial position and fortis plosives in word-medial position, there is a non-significant tendency for VOT to increase as the place of articulation goes back, in line with crosslinguistic tendencies [18], but the retroflexes /ʈ, ʈː/ always have the longest VOT across all places of articulation. As can be seen in the spectrograms in Figure 1, the magnitude of the positive VOT for retroflexes could be interpreted as affrication. Auditory impressions are that there is limited evidence for a sub-apical articulation for the retroflexes, but instead more of an apico-postalveolar articulation.

Taken together, these results indicate that an interpretation of the contrast between the Kufo plosives as 'fortis' vs. 'lenis', as proposed in [7], is reasonable. The nature of the contrast is similar to contrasts described as fortis-lenis in various Australian and Oto-Manguean languages, in that duration is a major but not the only correlate, and medial lenis plosives are prone to voicing and incomplete closures in medial environments [12][13], which in Kufo results in frequent approximant realisations. However, the acoustic phonetic results also highlight the complexities regarding conceptualisations of 'fortis-lenis' vs. 'geminate-singleton' contrasts, given the importance of duration as a cue in both cases, as well as various secondary cues. A crucial next step would be perceptual studies investigating which cues Kufo speakers are attending to, and whether duration is the most important. Articulatory studies of different sorts are also needed, in order to better understand the production differences between the two plosive series, as well as more comprehensive acoustic analyses, for example including intensity, f0, and burst amplitude. Future acoustic, articulatory and perceptual work on the sounds of Kufo would also ideally involve multiple speakers.

Although based on data collected with a single speaker, this study of Kufo plosives adds to our understanding of fortis-lenis contrasts in the world's languages. The acoustic evidence of the fortis-lenis contrast in Kufo pulmonic plosives shows that length and/or voicing contrasts established based on auditory impressions only may not always be accurate, and reinforces that contrasts between consonant series may involve multiple phonetic correlates rather than just one. The variation in voicing exhibited by Kufo lenis plosives explains why some early work suggests two pulmonic plosive series with a voicing contrast [3][4], whereas others suggest phonetic voicing conditioned by environment [5][6]. The clear role of duration as an important acoustic correlate distinguishing the fortis and lenis plosives in Kufo explains why some previous proposals include geminates [3][4]. This study highlights the need for more studies on consonant strength and length in African languages in order to add to our understanding of phonetic typology across the world's languages.

6. Acknowledgements

A sincere thank-you to Haroun Kafi for contributing his time and knowledge to this study, and to the two anonymous reviewers for their valuable feedback.

7. References

[1] Hammarström, H. & Forkel, R. & Haspelmath, M. & Bank, S., "Glottolog 5.0", Leipzig: Max Planck Institute for Evolutionary Anthropology, https://glottolog.org/resource/languoid/id/kang1288, 2024.

[2] Eberhard, D. M., Gary F. S., & Charles D. F., "Ethnologue: languages of the world", Dallas, Texas, SIL International, http://www.ethnologue.com, 2024.

[3] Hall, E. and Hall, M., "Kadugli-Krongo", Occasional Papers in the Study of Sudanese Languages, 9:57-67, 2004.

[4] Kafi, H. & Mongash, A., "Kufo alphabet book" [unpublished manuscript], 1998.

[5] Schadeberg, T. C., "Comparative Kadu wordlists", in Afrikanistische Arbeitspapiere: Schriftenreihe des Kölner Instituts für Afrikanistik 40: 11-48, 1994.

[6] Dafalla, R. Y., "A phonological comparison in the Katcha-Kadugli language group of the Nuba Mountains", in Insights into Nilo-Saharan Language, History and Culture: Proceedings of the 9th Nilo-Saharan Linguistic Colloquium, Institute of African and Asian Studies, University of Khartoum, 153-172, 2004.

[7] Li, S., "A phonological sketch of the Kufo language", Honours thesis, Australian National University, 2022.

[8] Ladefoged, P., & Maddieson, I., The Sounds of the World's Languages, Oxford: Blackwell Publishers, 1996.

[9] Cho, T. et al., "Acoustic and aerodynamic correlates of Korean stops and fricatives", in Journal of Phonetics 30(2):193-228, 2002.

[10] Kohler, K. J., "Phonetic explanation in phonology: the feature fortis/lenis", Phonetica 41(3):150-174 (1984).

[11] Jaeger. J., "The fortis/lenis question: evidence from Zapotec and Jawoñ", Journal of Phonetics, 11(2): 177-189, 1983.

[12] Stoakes, H., "An acoustic and aerodynamic analysis of consonant articulation in Bininj Gun-wok." PhD thesis, University of Mel-bourne, 2013.

[13] Dicanio, C., T., "The phonetics of fortis and lenis consonants in Itunyoso Trique", International Journal of American Linguistics 78(2):239-272, 2012.

[14] Boersma, P. and Weenick, D., "Praat: Doing phonetics by computer" [computer program], www.praat.org, 2016.

[15] Winkelmann, R., Harrington, J., & Jaensch, K., "EMU-SDMS: Advanced speech database management and analysis in R", Computer Speech & Language, 45, 392-410, 2017.

[16] R Core Team, "R: A language and environment for statistical computing" [computer program], https://www.R-project.org, Vienna, Austria, 2022.

[17] Winkelmann, R., Jaensch, K., Cassidy, S., & Harrington, J., "emuR: Main Package of the EMU Speech Database Management System", R package version 2.3.0, 2021.

[18] Nearey, T. M., & Rochet, B. L., "Effects of place of articulation and vowel context on VOT production and perception for French and English stops", Journal of the International Phonetic Association, 24(1), 1–18, 1994.

[19] Lisker, L. & Abramson, A. S., "A cross-language study of voicing in initial stops: Acoustical measurements", Word, 20(3), 384-422, 1964.

[20] Cho, T. & Ladefoged, P., "Variation and universals in VOT: evi-dence from 18 languages", Journal of Phonetics, 27, 207-229, 1999.

[21] Ladd, D. R. & Schmid, S., "Obstruent voicing effects on F0, but without voicing: Phonetic correlates of Swiss German lenis, fortis, and aspirated stops", in Journal of Phonetics, 71, 229-248, 2018.

[22] Burroni, F., Lau-Preechathammarach, R., & Maspong, S., "Unifying initial geminates and fortis consonants via laryngeal specification: Three case studies from Dunan, Pattani Malay, and Salentino", Proceedings of the Annual Meetings on Phonology, 2021.

[23] Bendjaballah, S., & Le Gac, D., "The acoustics of word-initial and word-internal voiced stops in Somali", Journal of the International Phonetic Association, 53(3), 644-681, 2023.

[24] Jessen, M., Phonetics and phonology of tense and lax obstruents in German. Amsterdam: John Benjamins, 1998.

On Consonant Gemination and Durational Effects in Word-Medial Position in Modern Standard Arabic: A Preliminary Study

Albandary Aldossari, Angelo Dian, John Hajek, Janet Fletcher

School of Languages and Linguistics, University of Melbourne

aaldossari@student.unimelb.edu.au; a.dian@unimelb.edu.au; j.hajek@unimelb.edu.au;
j.fletcher@unimelb.edu.au

Abstract

This study examines the short/long consonant contrast in word-medial position in Modern Standard Arabic (MSA) which also has contrastive vowel length. Experimental tokens containing Arabic singleton and geminate plosives (/b b:, t t:, k k:, d d:, tˁ tˁ:, q q:/) were recorded in a carrier sentence. Results confirm that geminates have significantly longer duration than singletons. The duration difference is smaller following long vowels. Preceding short vowels but not long vowels show some phonetic shortening before geminates. C/V1 ratio significantly increases with long /C:/ and decreases with long /V:/, with greater geminate-singleton differences following a short /V/.

Index Terms: gemination, duration ratio, Arabic, Speech Production

1. Introduction

1.1. Background

1.1.1. Stop gemination and acoustic correlates in MSA

Consonant length is fully phonemic in Arabic, which displays a phonological contrast between long (or geminate) and short (or singleton) consonants. This study examines the length contrast in Arabic stop sounds, /b b:, t t:, k k:, d d:, tˁ tˁ:, q q:/, as illustrated by the difference between /had:a/ 'demolished' and /hada:/ 'showed me the way.'

Consonant gemination has been an area of significant cross-linguistic interest in experimental phonetics ([1–5]), including Arabic. Most phonetic studies to date on gemination in Arabic (e.g., [3]; [6–7]; [8–12]) focus on colloquial varieties, e.g., Lebanese Arabic [13]. Colloquial Arabic can differ significantly in terms of phonetic and phonological form and behaviour according to where it is spoken as well as from Modern Standard Arabic (MSA) which is the standard variety used by educated speakers. There are very few acoustic studies on consonant gemination in MSA [3] and [6-7].

Previous acoustic-phonetic studies on Arabic (including MSA, e.g., [3] and [6-7]) have confirmed that closure duration is the primary acoustic correlate of gemination.

1.1.2. Interactions between vowel length and consonant gemination

Arabic is typologically unusual in that a phonemic length difference occurs not only for consonants but also for vowels, with both types of contrast also intersecting with each other: short and long vowels can appear before either short or long consonants, resulting in a 4-way contrast, e.g., /VCV/ v. /V:CV/ v. /VC:V/ v. /V:C:V/ in word-medial position.

Acoustic studies on gemination have extended beyond the geminate or singleton consonants themselves to examine the interactions between vowel duration on the one hand and adjacent consonant gemination on the other (e.g., [13]; [14]). [15] notes that across languages (e.g., Finnish) geminates are generally associated with a shortening of the preceding vowel, while singletons are usually accompanied by increased duration of the preceding vowel. The languages he refers to involve in almost all cases research results only for phonologically short vowels. However, he cites data for Tamil which like Arabic also has a 4-way contrast to show that vowels in that language, regardless of length category, are shorter before a long consonant - albeit with a bigger effect noted for /V:C:/ vs. /V:C/ (32.5 ms reduction in the former) than for /VC:/ vs. /VC/ (14 ms difference).

In Arabic, vowels preceding geminate consonants are also reported to be subject to some phonetic shortening ([3], [6–7], [10], [13], [16]). However, almost all these studies focus only on the short vowel category before long and short consonants, i.e., /VC/ and /VC:/. [16] for instance, found for Iraqi Arabic a phonetic shortening effect (-16 ms) before geminate consonants (76ms) when compared to the pre-singleton position (92 ms). However, in a rare study to consider the interaction between the consonant and vowel length contrasts, [9] found for Lebanese Arabic that the duration of short vowels is the same before short and long consonants. Instead, there was a small but significant shortening of long vowels before geminates (-17 ms). With respect to C length and duration, there was no effect of vowel length on /C:/. However, short /C/ was slightly longer (+14 ms) after long /V/. In the only previous study to look at both consonant and vowel length in MSA, [7] reported only a small durational difference between short vowels (-5 ms), and of short /C/ after long v. short /V:/ (+6 ms). He also noted phonetically shorter /V:/ (-13 ms) before geminates

One additional useful way to consider gemination is the use of C/V duration ratio values. Previous research on Arabic has focussed on CC/C ratio to establish the relative duration of long to short consonants (e.g., [9]). One element that has generally not been explored previously with respect to word-medial gemination in Arabic (but see [17] on heteromorphemic gemination in Moroccan Arabic) is C/V ratio, which has been used in the analysis of gemination in other languages (e.g. Italian [18, 19], Japanese [20]). The C/V ratio has been regarded as a useful relational measure to normalise for variation in speech and has been found to be a particularly reliable correlate of gemination in languages that show a complementary relationship between the duration of consonants and preceding vowels in the V-C interval, such as Italian [19]. This potential relationship and what it means for MSA remains unexplored.

1.2. Aims

The primary aim of this study is to provide quantitative experimental acoustic findings on the duration of geminate consonants in word-medial positions in MSA. Specifically, it investigates whether geminate consonants are significantly longer than singletons, as reported for other regional Arabic varieties. Additionally, the study explores the effect of the length of the preceding vowel on MSA word-medial gemination, and how consonant length impacts the phonetic duration of short and long vowels. We are particularly interested in whether: (a) there is a concomitant decrease in vowel duration before geminate consonants, and (b) the extent to which the C/V1 ratio increases due to gemination across different vowel length categories.

Unlike previous research specifically on MSA, e.g., [7], who in a study published in Arabic analyzed geminates in word-medial positions with a focus on voicing status, our study focusses instead on the C/V1 ratio to quantify the relationship between geminate consonants and preceding vowels in MSA, offering new insights.

Based on most previous findings, we predict that short vowels will exhibit phonetic shortening before geminate consonants in MSA, as seen in previous studies on Tamil and MSA (e.g., [3], [6-7]) and Arabic dialects (e.g., [10], [13], [16]). However, we expect long vowels to show less or no shortening due to mixed findings for Arabic. Additionally, we predict that geminate consonants (/C:/) will have significantly longer durations than singleton consonants (/C/), consistent with prior findings on MSA [6-7] and other Arabic varieties (e.g., [9], [16]). Finally, we anticipate that the C/V ratio will show a noticeable rise with consonant gemination across both short and long vowel categories. Given that this ratio highlights the balance between consonant and vowel durations within the V-C interval, we expect the impact of gemination to be particularly striking when the vowel is categorically short, as the increased consonantal duration will play a more dominant role in this context.

2. Methods

2.1. Participants

Three Arabic native speakers (2 male and 1 female) took part in data collection. Their ages ranged from 21 to 39. All participants were university students or university educated. They were all MSA speakers, born and raised in Saudi Arabia. At the time of recording, all participants were present in Melbourne for academic pursuits. The participants were not reported to have any history of speech or language impairments.

2.2. Materials and procedure

The original experiment included 180 target words, featuring either a medial singleton or geminate consonant across a series of manners of articulation. We only consider stops here. Some words are minimal pairs, while others are near-minimal pairs, with geminate and singleton consonants following a stressed vowel (e.g., /ˈhaba:/ ‘escaped’ vs. /ˈhab:a/ ‘blew’).

Specifically, the experiment selected two sub-sets of disyllabic Arabic words, each illustrating distinct phenomena: a) the contrast between singleton vs. geminate consonants and b) contrastive vowel length before singleton vs. geminate consonants. This allowed for the 4-way length distribution, i.e., /VC/ /VC:V/, /V:CV/ /V:C:V/ in Arabic to be tested with a geminate. Below are instances of the two groups.

A diverse set of Arabic words, including 12 different stop phonemes categorized by place and voice voicing (voiceless: /k-k:, t-t:, tˤ-tˤ:, q-q:/ and voiced: /b-b:, d-d:/) in word-medial position, were carefully selected and examined in this study. Each stop was preceded by the short or long low central vowel, i.e. /a/ or /a:/, and followed by /a:/ or the short high back vowel /u/, as detailed in Table 1.

Two pairs of words were included for each target singleton and geminate sound. All target words in the experiment were disyllabic, with stress placement always on the first syllable.

All participants were asked to read MSA sentences with target words containing one of the geminate-singleton pair words, e.g., /ana: aqu:l (the target word) θa:ni:ja/ < ... نأا قاول ناثية> “I say (the target word) again.” Geminate diacritics in Arabic were marked in the target words and in the task (e.g., ّ<ب>) to avoid potential confusion.

The participants were presented with a Microsoft PowerPoint (PPT) presentation, with each slide containing sentences (including target words – geminate and singleton). The sentence presentation order was randomized. Each sentence was read 5 times. The total number of tokens, after a small number were excluded as errors, utilized in the study for the three speakers amounted to 1075 tokens.

The data were recorded in the Horwood Recording Studio at the University of Melbourne using a Charter Oak E700 dual diaphragm solid state condenser microphone and a Focusrite Scarlett 18i20 gen3 recording interface. The recordings were made at a sampling rate of 44.1 kHz and a quantization rate of 16 bits.

Table 1. *Sample Experimental Word List for Stop Sounds.*

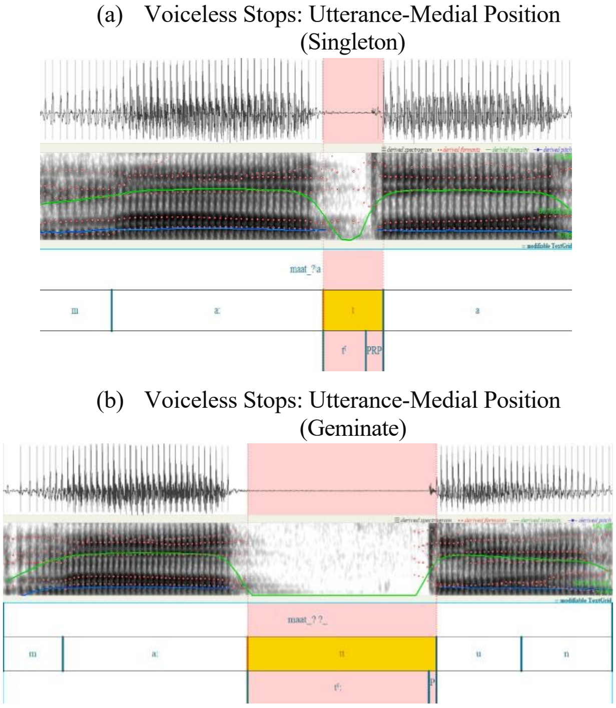
	Ph	Sing CVCV:	Gem CVC:V	Sing CV:CV	Gem CV:C:un
Voiceless Stops	/t/	/mata:/ ‘when’	/mat:a/ ‘Related by kinship’	/ma:ta/ ‘died’	/ma:t:un/ ‘formed a bond with someone’
	/k/	/ʃaka:/ ‘complained’	/ʃak:a/ ‘doubted’	/ʃa:ka/ ‘complained’	/ʃa:k:un/ ‘doubting’
	/q/	/saqa:/ ‘irrigated/ to water’	/ʃaq:a/ ‘caused to split/tear’	/sa:qa/ ‘narrated (the story)’	/ʕa:q:un/ ‘a disobedient child/boy’
	/tˤ/	/χatˤa:/ ‘walked’	/χatˤ:a/ ‘drew/sketched’	/χa:tˤa/ ‘sewed’	/χa:tˤ:un/ ‘drawer’
Voiced Stops	/b/	/ħaba:/ ‘crawled’	/hab:a/ ‘(The wind) blew’	/ʃa:ba/ ‘Grayed/aged’	/ħa:b:un/ ‘a loving person’
	/d/	/hada:/ ‘showed me the way’	/had:a/ ‘demolished’	/sˤa:da/ ‘hunted’	/ha:d:un/ ‘destroying’

2.3. Acoustic analysis

In the analysis for this study, force-aligned annotations of the recordings were initially conducted using WebMAUS [21] to automatically segment and align the speech data, followed by careful manual adjustments to refine the boundaries of consonant (C) and preceding vowel (V1) segments in PRAAT [22]. An additional tier ‘Phonetic’ was added to annotate closure and release characteristics of voiceless and voiced stops. Voiceless stops, shown in Figure (1) (a and b), were annotated as follows. The closure phase was identified from the point where the F1 and F2 pattern of the preceding vowel ended and continued until the release of the stop including the stop burst and any aspiration, referred to here as the post-release phase (PRP). The PRP was specifically defined as the time from the release of the stop to the onset of the first formant of the subsequent vowel. For the purposes of this study, whole consonant duration is considered (closure + PRP).

Voiced stops were analyzed with a method similar to voiceless stops to determine the onset of closure and the consonant itself. Unlike voiceless stops, however, voiced stops usually showed a sharp decrease in F1 and F2 energy with low energy voicing throughout the closure phase i.e., a voice bar. The offset of the consonant was identified by a sharp increase in amplitude and energy in F2 and higher frequencies after the release phase.

Consonant duration was measured by calculating the interval between the onset of closure and the onset of the following vowel. V1 duration was measured by calculating interval between the onset and offset of F1 and F2 energy associated with the vowel as indicated in Figure 1a.


(a) Voiceless Stops: Utterance-Medial Position (Singleton)

(b) Voiceless Stops: Utterance-Medial Position (Geminate)

Figure 1: *Examples of two annotated utterance-medial target words produced by an Arabic male speaker, indicating overall C interval plus (1) closure phase for singletons (a) and geminates (b), (2) post-release phase ([PRP]), and (3) preceding vowel intervals*

2.4. Statistical analysis

Data extraction was performed in R [23] using the 'emuR' package that is part of the emu-SDMS software suite [24]. Specifically, duration values were extracted for medial voiced and voiceless stops and the preceding vowel (V1) for all experimental tokens. In this study only overall C duration (closure + PRP interval) was measured. Three separate Linear Mixed-effects Models (LMMs) implemented via the 'lmerTest' package in R [25], examined the effects and interactions of C gemination and V1 phonological length on three key phonetic parameters: C duration, V1 duration, and the ratio of C duration to V1 duration (C/V1 ratio) for all speakers. These models incorporated fixed effects for gemination (singleton vs. geminate) and V1 length (short vs. long). Random intercepts and slopes were included for *Speaker* and *Word* and random slopes for gemination by *Speaker*. Model simplification was also employed using 'step.' Post-hoc analyses were conducted using 'emmeans' to further explore significant interactions among the fixed factors.

3. Results

3.1. C duration

Table 2 and Figure 2 report mean C duration depending on whether C is a singleton or geminate and whether the preceding vowel (V1) is long or short. There is a highly significant main effect of gemination ($F(1,8.97) = 825.304$, $p < .0001$). All three speakers produce significantly longer geminate stops compared to singleton stops. There is also a significant interaction between C and V1 length categories ($F(1,68.61) = 7.397$, $p < .01$). Post-hoc tests revealed a somewhat reduced geminate-singleton C duration difference following a long vowel (β = 142 ms, $p < .001$) as compared to a short vowel (β = 165 ms, $p < .001$).

Table 2. *Mean C duration (ms), standard deviation (SD), and token counts (N) by V1 and C lengths with speakers pooled.*

V1 length	C length	Mean C dur (ms)	SD	N
long	gem	262	41	269
long	sing	121	22	269
short	gem	279	30	270
short	sing	115	23	267

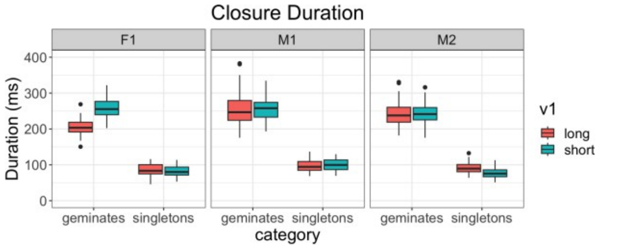


Figure 2: *Consonant duration (ms) plotted by Consonant length category and preceding vowel length category for the three speakers.*

3.2. V1 duration

Table 3 and Figure 3 present the duration of V1 (long or short) preceding either geminates or singletons. The overall effect of C length was not found to be significant ($F(1,68.33) = 1.07$, $p = 0.3167$), indicating that V1 duration does not vary with gemination. Conversely, there is a highly significant main effect of V1 length ($F(1,2.22) = 272.352$, $p < .01$), with phonologically long vowels always longer on average than phonologically short vowels. There is also a significant interaction between gemination and V1 length categories ($F(1,68.33) = 10.613$, $p < .01$). Post-hoc tests reveal that V1 is only ~10 ms shorter preceding geminates than singletons (β = -11 ms, $p < .01$) when it is phonologically short, while there is no statistically significant difference when it is phonologically long.

Table 3. *Mean V1 duration (ms), standard deviation (SD), and token counts (N) by V1 and C lengths with speakers pooled.*

V1 length	C length	Mean V1 dur (ms)	SD	N
long	gem	243	36	269
long	sing	237	33	269
short	gem	77	20	270
short	sing	87	18	267

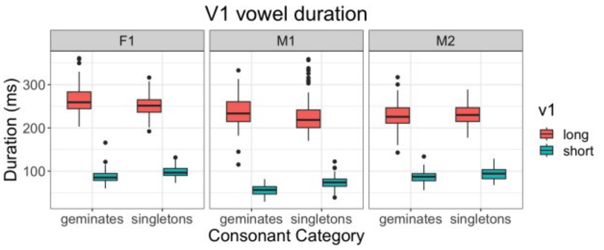


Figure 3: *Vowel duration (ms) plotted according to phonological vowel length and following consonant length category for the three speakers.*

3.3. C/V1 ratio

Table 4 and Figure 4 illustrate variation in the C/V1 ratio according to C length and V1 length. There is a significant main effect of C length (*F*(1,2.5) = 39.025, p < .05), with higher C/V1 ratio for geminates than singletons across V1 length types. There is also a significant main effect of V1 length (*F*(1,2.14) = 18.15, p < .05), with increased C/V1 ratio for short relative to long V1, regardless of C length. The interaction between consonant and vowel length is highly significant (*F*(1,68.53) = 122.647, p < .0001), suggesting that the impact of consonant length on the ratio varies considerably with vowel length. A post-hoc analysis showed a reduced difference in the C/V1 ratio between long and short consonants when the preceding vowel was long (β = 0.59) compared to when it was short (β = 2.54).

Table 4: *Mean C/V1 ratio, standard deviation (SD), and token counts (N) by V1 and C lengths with speakers pooled.*

V1 length	C length	Mean C/V1 ratio	SD	N
long	gem	1.11	0.27	269
long	sing	0.52	0.13	269
short	gem	4.00	1.51	270
short	sing	1.2	0.61	267

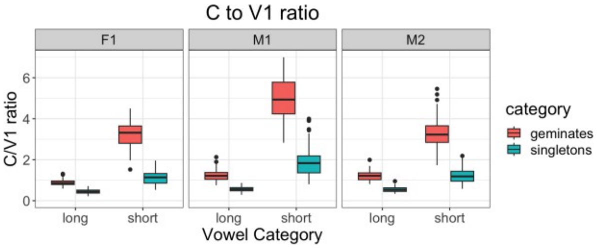


Figure 4: *C/V1 plotted according to phonological vowel length and following consonant length category for the three speakers.*

4. Discussion and Conclusion

In the first instance, this study aimed to determine whether the duration of geminate consonants was longer than that of singletons in medial word positions in MSA. Furthermore, the study investigated (a) whether the phonetic duration of the preceding vowel also co-varied with consonant length for both long and short vowels and (b) whether the C/V1 ratio can be identified as a useful acoustic measure of gemination in MSA. The results have shown that geminates have longer durations than singletons for all speakers, consistent with previous research on different varieties of Arabic (including MSA as well as on other languages, confirming closure duration as the primary acoustic correlate for gemination).

Another point to note is that although this study does not find a significant overall effect of C gemination on the duration of the preceding vowel (V1), in the specific case of phonologically short V1, there is a small but significant shortening effect (~10 ms) preceding geminates. This is in line with cross-linguistic tendencies of pre-geminate vowel shortening [15] which have also previously been reported for Arabic short vowels (e.g. [16]), although not consistently (cf. §1.1.2). However, this pre-geminate shortening effect is not observed for long vowels in this study, in contrast with Lebanese Arabic [9] whereby long but not short V1 shortens before geminates. Similarly, it does not fully align with previous results for MSA [7] which found a bigger phonetic effect on long vowels before geminates.

Furthermore, this study finds that the C/V1 ratio is significantly affected by both C and V1 lengths in MSA. However, a markedly greater geminate-singleton difference is observed following short vowels as compared to long vowels. This is unsurprising as it merely suggests that the effect of C gemination is more visible on C duration where the relative contribution of preceding V1 duration in the V1-C interval is smaller due to its categorically short duration. This differs from Italian whereby the durations of V1 and C are negatively correlated (i.e., the duration of V1 decreases linearly with longer C duration as well as categorically with C length) and the C/V1 ratio provides a measure of this relationship [18]. However, from a normalization perspective, as found for other languages, the C/V ratio may serve as a more stable correlate of gemination than absolute duration across speaking rates in MSA as well as other Arabic varieties, although this needs to be tested in future investigations specifically considering variation in speaking rate.

Overall, the findings indicate that duration patterns play a key role in differentiating between geminate and singleton consonants in MSA. This is anticipated in a language that differentiates phonemic consonant (and vowel) lengths, highlighting the significance of timing in length distinctions.

However, it is important to acknowledge the limitations of this analysis, which was based on data from three speakers and focused on a limited dataset. Future research will broaden the scope to include factors like word-medial and word-final positions, different manners of articulation, phonological voicing, and the impact of both preceding and following vowels of the target words. Including a larger sample of speakers will also help validate the results across different manners of articulation and voicing statuses.

5. References

[1] A. Lahiri and J. Hankamer, "The timing of geminate consonants," *Journal of Phonetics*, vol. 16, pp. 327-338, 1988. doi: 10.1016/S0095-4470(19)30506-6.

[2] W. Ham, *Phonetic and phonological aspects of geminate timing*. New York, NY: Routledge, 2001.

[3] M. Y. Frej, "The production and perception of peripheral geminate/singleton coronal stop contrasts in Arabic," Ph.D. dissertation, Western Sydney Univ., Sydney, Australia, 2021.

[4] J. Al-Tamimi and G. Khattab, "Acoustic correlates of the voicing contrast in Lebanese Arabic singleton and geminate stops," *Journal of Phonetics*, vol. 71, pp. 306-325, 2018.

[5] J. Al-Tamimi and G. Khattab, "Multiple cues for the singleton-geminate contrast in Lebanese Arabic: Acoustic investigation of stops and fricatives," in *Proc. 17th Int. Congr. Phonetic Sciences*, Hong Kong, China, 2011, pp. 212-215.

[6] K. Ferrat and M. Guerti, "An experimental study of the gemination in Arabic language," *Archives of Acoustics*, vol. 42, no. 4, pp. 571-578, 2017.

[7] Y. A. Ahmad, "The effect of gemination on vowel length in Arabic," *Arab Journal for the Humanities*, vol. 117, pp. 11-47, 2012. (in Arabic).

[8] J. Al-Tamimi and G. Khattab, "Acoustic cue weighting in the singleton vs geminate contrast in Lebanese Arabic: The case of fricative consonants," *The Journal of the Acoustical Society of America*, vol. 138, no. 1, pp. 344-360, 2015.

[9] G. Khattab and J. Al-Tamimi, "Geminate timing in Lebanese Arabic: The relationship between phonetic timing and phonological structure," *Laboratory Phonology*, vol. 5, no. 2, pp. 231-269, 2014.

[10] M. Al-Deaibes and N. Rosen, "Gemination in Rural Jordanian Arabic," in *Perspectives on Arabic Linguistics XXX*, 2019, pp. 53-76.

[11] A. Issa, "Durational and non-durational correlates of lexical and derived geminates in Arabic," in *Proc. INTERSPEECH 2023*, Dublin, Ireland, 2023, pp. 4753-4757. doi: 10.21437/Interspeech.2023-2187.

[12] A. G. E. Issa, "Phonetic and Phonological Aspects of Gemination in Libyan Arabic," Ph.D. dissertation, Univ. of Leeds, Leeds, UK, 2016.

[13] G. Khattab, "A phonetic study of gemination in Lebanese Arabic," in *Proc. ICPhS XVI*, Saarbrücken, Germany, 2007, pp. 153-158.

[14] C. S. Doty, K. Idemaru, and S. G. Guion, "Singleton and geminate stops in Finnish - Acoustic correlates," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007, pp. 2737-2740.

[15] I. Maddieson, "Phonetic cues to syllabification," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, V. Fromkin, Ed. New York, NY: Academic Press, 1985, pp. 203-221.

[16] Z. M. Hassan, "Gemination in Swedish and Arabic with a particular reference to the FONETIK 2002," in *Proc. FONETIK 2002*, Stockholm, Sweden, 2002, pp. 81-84.

[17] R. Ridouane and G. Turco, "Why is gemination contrast prevalently binary? Insights from Moroccan Arabic," *Radical: A Journal of Phonology*, vol. 1, pp. 62-91, 2019.

[18] A. Dian, J. Hajek, and J. Fletcher, "Cross-regional patterns of obstruent voicing and gemination: The case of Roman and Veneto Italian," *Languages*, under review.

[19] E. R. Pickett, S. E. Blumstein, and M. W. Burton, "Effects of speaking rate on the singleton/geminate consonant contrast in Italian," *Phonetica*, vol. 56, pp. 135–157, 1999.

[20] Y. Hirata and J. Whiton, "Effects of speaking rate on the singleton/geminate distinction in Japanese," *Journal of the Acoustical Society of America*, vol. 118, pp. 1647–1660, 2005.

[21] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[22] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer* (Version 6.0.19), 2016. Available: http://www.praat.org/.

[23] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2003. Available: https://www.R-project.org.

[24] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, vol. 45, pp. 392-410, 2017.

[25] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1-26, 2017.

Perception of the Kelantan Malay Word-Initial Singleton/Geminate Contrast by Vietnamese Speakers

Mohd Hilmi Hamzah[1], *Kimiko Tsukada*[2,3], *John Hajek*[3], *Đích Mục Đào*[4]

[1]Universiti Utara Malaysia, Malaysia, [2]Macquarie University, Australia, [3]The University of Melbourne, Australia, [4]University of Social Sciences and Humanities, Vietnam National University - Ho Chi Minh City, Vietnam

hilmihamzah@uum.edu.my, kimiko.tsukada@gmail.com, j.hajek@unimelb.edu.au, dich.daovns@gmail.com

Abstract

This study investigates the perception of the Kelantan Malay (KM) word-initial consonant length contrast among Vietnamese speakers who are naïve to such a contrast. One Vietnamese group and a native KM control group participated in the AXB discrimination experiment. As expected, KM listeners outperformed Vietnamese listeners in discriminating KM consonant length, with the overall discrimination accuracy being 88% and 53% for the KM and Vietnamese groups, respectively. While there was a clear between-group difference in discrimination accuracy, there was also greater variability among the KM group, suggesting genuine difficulty in successfully identifying word-initial geminates, particularly those beginning with voiceless stops.

Index Terms: speech perception, consonant length contrast, singleton, geminate, Kelantan Malay, Vietnamese

1. Introduction

In KM, length (i.e., short vs long) is lexically contrastive across all consonants, as empirically established in [1-4]. For example, /kabo/ 'beetle' contrasts word-initially with /kkabo/ 'blurry'. It has been reported that closure duration is the most robust acoustic correlate of the singleton/geminate contrast in KM, with non-durational parameters such as root mean square (RMS) amplitude and fundamental frequency (F0) playing important secondary roles in enhancing the word-initial length contrast in this language. Consonant length in KM is claimed to be cross-linguistically rare and more marked as it only occurs in word-initial position [e.g., 5]. It is well known that the word-initial length contrast is more perceptually indiscernible, particularly for utterance-initial geminates involving voiceless stops, such as /kk/ in /kkabo/. In this particular utterance position, there is ostensibly insufficient acoustic information available for listeners to discern the length contrast [e.g., 6].

Unlike KM, consonant length is not contrastive in Vietnamese [e.g., 7-9]. Vietnamese speakers, however, are familiar with contrastive duration of vowels that is employed to distinguish some tense-lax vowels in the Vietnamese language [10]. It is thus of theoretical interest to examine whether Vietnamese listeners, whose L1 uses duration differently, might adopt varied perceptual strategies when faced with tasks involving consonant length contrasts in a non-native language, such as KM in this study. This is evident in a related study on cross-language perception of phonological length contrasts by [11] who compared German and Italian sounds. It was found that German speakers, who have experience with vowel length contrasts in their L1, were less accurate in distinguishing consonant length in Italian. Conversely, Italian speakers were as proficient as German speakers in discerning vowel length in German.

With regard to Vietnamese speakers, it has been reported in [12] that it was problematic for Vietnamese listeners to perceive word-medial consonant length in Japanese. This was supported by [7] and [13] who found that Vietnamese speakers had a tendency to misidentify singletons and geminates in Japanese. As for KM speakers, who are familiar with consonant length in their L1, it was reported in [14] that they are able to perceive Japanese consonant length at a high level of discrimination accuracy. Similarly, speakers from other languages with L1 contrastive length, such as Korean and Mongolian speakers [15], are also able to perceive Japanese consonant length above chance level. However, those who lack experience in consonant length are less accurate in their perception of Japanese consonant length, such as American English [16] and Mandarin [15] speakers.

In this study, the perception of KM consonant length by native and non-native speakers, i.e., Vietnamese speakers, was compared to examine the extent to which word-initial consonant gemination in KM is processed accurately by speakers of other languages who are unfamiliar with consonant length contrasts. We hypothesise that Vietnamese speakers may have difficulties in perceiving the unfamiliar, rare linguistic event of word-initial consonant gemination in KM. Given the complexities of the word-initial length contrast, we also expect potential variation in discrimination accuracies among KM native listeners due to effects of voicing and manner of articulation. Our findings will provide some theoretical insights into this potential cross-linguistic phonetic transfer and add to our current knowledge about the perception of difficult sounds such as word-initial consonant gemination.

2. Methods

2.1. Stimuli preparation

2.1.1. *Speakers and procedures*

The experimental stimuli were obtained from previous production experiments in KM [3, 4] involving sixteen native speakers (8 males, 8 females). Six of them were students from several universities in Melbourne, Australia, and ten were students from Universiti Malaysia Kelantan located in the state of Kelantan, Malaysia. Their age ranged from 20 to 28 (mean age: 22.4). All were born and raised in Kelantan, Malaysia. For the speakers in Melbourne, the experimental materials were recorded individually on the main campus of the University of Melbourne. As for the speakers in Kelantan, they were recorded individually in a quiet room at Universiti Malaysia Kelantan.

In all sessions, speakers were asked to repeat each token in isolation and in a carrier sentence. The carrier sentence was: /diɔ katɔ (the target word) tigɔ kali/ "he said (the target word) three times", adapted from [17]. All experimental tokens were presented in randomised order using a powerpoint presentation on a computer. The experiment took approximately one and a half hours for each speaker. They were compensated financially for their participation in the experiment.

In the present study, the production data collected in Kelantan, Malaysia were selected as experimental stimuli, which involved six native speakers of KM (3 males, 3 females). Only tokens produced in isolation, i.e., utterance-initial position, were used.

2.1.2. Speech materials

Table 1. *Nineteen pairs of Kelantan Malay words with target sounds underlined and bolded.*

Phoneme pair	Singleton		Geminate	
	Word	Gloss	Word	Gloss
/p/–/pp/	/**p**itu/	door	/**pp**itu/	at the door
	/**p**agi/	morning	/**pp**agi/	early morning
/t/–/tt/	/**t**ido/	sleep	/**tt**ido/	sleep by chance
	/**t**anɔh/	land	/**tt**anɔh/	outside
/k/–/kk/	/**k**iɣi/	left	/**kk**iɣi/	to the left
	/**k**abo/	blurry	/**kk**abo/	beetle
/b/–/bb/	/**b**ini/	wife	/**bb**ini/	married
	/**b**atʃɔ/	read	/**bb**atʃɔ/	is reading
/d/–/dd/	/**d**ike/	song	/**dd**ike/	sing a song
	/**d**apo/	kitchen	/**dd**apo/	at the kitchen
/g/–/gg/	/**g**igi/	teeth	/**gg**igi/	on the teeth
	/**g**adʒi/	salary	/**gg**adʒi/	sawing tool
/m/–/mm/	/**m**isa/	moustache	/**mm**isa/	moustached
	/**m**aɣi/	come	/**mm**aɣi/	cupboard
/n/–/nn/	/**n**ikɔh/	marriage	/**nn**ikɔh/	married
	/**n**anɔh/	pus	/**nn**anɔh/	getting pus
/ŋ/–/ŋŋ/	/**ŋ**aŋɔ/	open the mouth	/**ŋŋ**aŋɔ/	agape
/l/–/ll/	/**l**idɔh/	tongue	/**ll**idɔh/	on the tongue
	/**l**apu/	lights	/**ll**apu/	on the lights

Table 1 shows nineteen KM word pairs used in this study. All tokens were disyllabic words with either C(C)VCV or C(C)VCVC structures. They contained singletons (n = 114) or geminates (n = 114) in word-initial position (underlined and bolded). These phonemes were grouped according to consonant type: voiceless stops (/p/–/pp/, /t/–/tt/, /k/–/kk/); voiced stops (/b/–/bb/, /d/–/dd/, /g/–/gg/); and sonorants consisting of nasals (/m/–/mm/, /n/–/nn/, /ŋ/–/ŋŋ/) and liquids (/l/–/ll/). Each phoneme was followed by two distinct vowels: the high front vowel /i/ and the low central vowel /a/, except /ŋ/–/ŋŋ/ (low central vowel /a/ only).

On average, the VOT durations for voiceless stops were 29 ms and 18 ms for singletons and geminates, respectively. As for voiced stops, the closure durations were 60 ms for singletons and 152 ms for geminates (geminate-to-singleton ratio: 2.51). For sonorants, the closure durations were 58 ms and 158 ms for singletons and geminates, respectively (geminate-to-singleton ratio: 2.69). These durational values are in good agreement with what has been reported in previous research [e.g., 1, 3].

2.2. Participants

Two groups of young adults participated in an AXB discrimination task. The first group consisted of 24 (10 males, 14 females) native speakers of Vietnamese (mean age = 27.1). They were initially divided into two sub-groups, with the first one involving either academics or students at Vietnam National University, Ho Chi Minh City, Vietnam, while the second one involved general workers outside the university setting. Given their similar results (see below), we decided to merge them together. All of them were born and raised in Vietnam and are fluent in Vietnamese. All were naïve to KM.

The second and a control group consisted of 12 (6 males, 6 females) native speakers of KM (mean age = 39.5). They were either academic or non-academic staff members at Universiti Utara Malaysia in Kedah, Malaysia. All KM participants were born and spent the majority of their life in Kelantan, Malaysia. None of them participated in the recording sessions. According to self-report, all had normal hearing at the time of the experimental sessions.

2.3. Procedure

For the Vietnamese participants, they were tested individually in a quiet room at their workplace or on the university campus of Vietnam National University. As for the KM participants, the experiments were conducted in a quiet room on the main campus of Universiti Utara Malaysia in Sintok.

In all sessions, the experiment was self-paced and lasted approximately 15 to 20 minutes. The participants heard the stimuli at a self-selected, comfortable amplitude level over the high-quality speakers on a notebook computer. They completed a two-alternative forced-choice AXB discrimination task, in which they were asked to listen to trials arranged in a triad (A-X-B). The presentation of the stimuli and the collection of perception data were controlled by the PRAAT program [18]. In the AXB task, the first (A) and third (B) tokens always came from different length categories, and the participants had to decide whether the second token (X) belonged to the same category as A (e.g., 'pitu$_2$'-'pitu$_1$'-'ppitu$_3$') or B (e.g., 'kabo$_3$'-'kkabo$_1$'-'kkabo$_2$'; where the subscripts indicate different speakers).

The participants listened to a total of 160 unique trials. The first eight trials were for practice and were not analysed. The three tokens in all trials were spoken by three different speakers. Thus, X was never acoustically identical to either A or B. This was to ensure that the participants focused on relevant phonetic characteristics that grouped two tokens as members of the same length category without being distracted by audible but phonetically irrelevant within-category variation (e.g., in voice quality). This was considered a reasonable measure of participants' perceptual capabilities in real world situations [19]. All possible AB combinations (i.e., AAB, ABB, BAA, and BBA, 38 trials each) were tested.

The participants were given two ('A', 'B') response choices on the computer screen. They were asked to select the option 'A' if they thought that the first two tokens in the AXB sequence were the same and to select the option 'B' if they thought that the last two tokens were the same. No feedback was provided during the experimental sessions. The participants could take a break after every 40 trials if they wished. The participants were required to respond to each trial, and they were told to guess if uncertain. A trial could be replayed as many times as the participants wished in order to reduce their anxiety, but responses could not be changed once given. The interstimulus interval in all trials was 0.5 s.

3. Results

We used R version 4.4.0 for statistical analyses and data visualisation reported below [20]. The packages used include ez [21] and tidyverse [22].

3.1. Overall results

Figure 1 shows the distributions of percentages of correct discrimination by the two groups of participants. The overall mean discrimination accuracy was clearly higher for the native KM control group (88%) than the Vietnamese group (53%). Note that the accuracy score for the Vietnamese group was just above the 50% chance-level. A comparison via the Welch two-sample *t*-test showed that the difference between the KM and Vietnamese groups was significant [$t(15.3) = 34$, $p < .001$]. Nonetheless, as seen in Figure 1, there is greater individual variation in the discrimination accuracies among the KM participants (*sd*=0.1) as compared to the Vietnamese participants (*sd*=0.04). The highest discrimination accuracy for the KM group was 97%, while the lowest was 65%, indicating that the discrimination task was somewhat challenging for some native speakers. As for the Vietnamese group, the discrimination accuracy ranged from 47% to 63%, with six of the 24 scoring below the 50% chance-level.

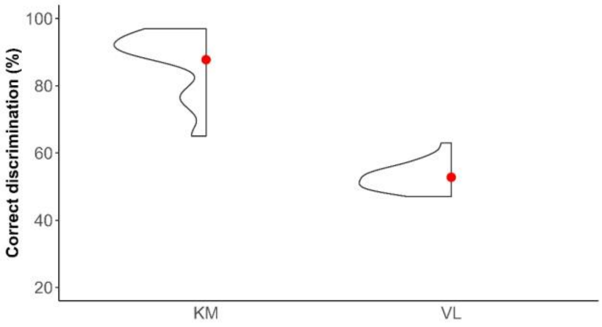


Figure 1: *Accuracy (%) of length discrimination by two groups of participants (KM: Kelantan Malay; VL: Vietnamese). The red circle indicates the mean.*

3.2. Comparison of the length category (Geminate vs Singleton) of the target token (X in AXB)

Figure 2 shows the distributions of percentages of correct discrimination for trials differing in the length category (geminate, singleton) of the target token. The question of interest was if the participants' discrimination accuracy differed between trials in which X in AXB was a geminate and trials in which X in AXB was a singleton. Two-way analysis of variance (ANOVA) results with group (KM, Vietnamese) and length category (geminate, singleton) reached significance only for the main effect of group [$F(1, 68) = 391.7$, $p < .001$, $\eta^2 = .85$]. That is, the KM group was significantly more accurate than the Vietnamese group whether X in the AXB sequence was singleton or geminate. Length category did not yield any significant main effect on discrimination accuracies. Two-way interaction also yielded a non-significance result. As seen in Figure 2, neither group was biased with respect to the length category of the target token, though the Vietnamese group shows a higher discrimination accuracy for geminates (55%) than for singletons (51%). As for the KM group, the discrimination accuracies for geminates and singletons were both similar (88% for both geminates and singletons).

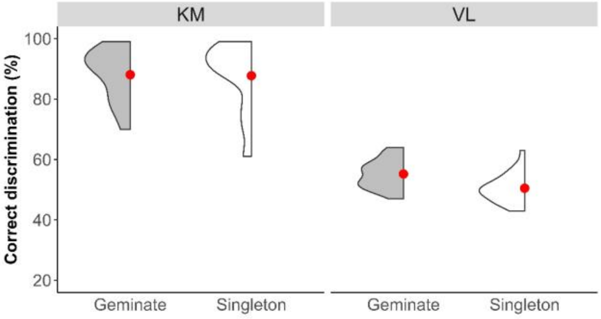


Figure 2: *Accuracy (%) of length discrimination for trials differing in the length category of the target token.*

3.3. Comparison of the consonant type (Voiceless stop vs Voiced stop vs Sonorant) of the target token (X in AXB)

Figure 3 shows the distributions of percentages of correct discrimination for trials differing in consonant type (voiceless stop, voiced stop, sonorant) of the target token. The question of interest was to determine if the participants' discrimination accuracy differed between trials in which X in AXB was a voiceless stop, a voiced stop, or a sonorant. Two-way ANOVA results with group (KM, Vietnamese) and consonant type (voiceless stop, voiced stop, sonorant) reached significance for the main effects of group [$F(1, 354) = 658.3$, $p < .001$, $\eta^2 = .65$] and also consonant type [$F(2, 354) = 3.7$, $p < .05$, $\eta^2 = .02$]. Two-way interaction yielded a non-significance result.

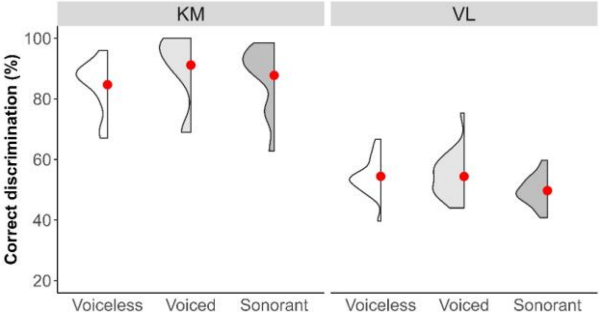


Figure 3: *Accuracy (%) of length discrimination for trials differing in the consonant type of the target token.*

As observed in Figure 3, both groups of participants appeared to be biased with respect to the consonant type of the target token. Note that the order of discrimination accuracy (from the lowest to the highest) was different between the two groups. On one hand, for the KM group, voiceless stops received the lowest discrimination accuracy (85%), while voiced stops were best discriminated (91%). Sonorants were intermediate between the other consonant types (88%). On the other hand, for the Vietnamese group, sonorants were discriminated with the lowest discrimination accuracy (50%) as compared to voiceless stops and voiced stops (both 54%). The comparison of phonemes within these consonant types is described in the following section.

3.4. Comparison of the phonemes of the target token (X in AXB)

Figures 4 and 5 below show the distributions of percentages of correct discrimination for trials differing in the phonemes of the target token for the KM and Vietnamese groups, respectively. The phonemes in these figures are arranged from the lowest (left) to the highest (right) discrimination accuracies. As seen in Figure 4 for the KM group, the voiceless alveolar stop pair (/t/-/tt/) was discriminated with the lowest discrimination accuracy (83%), while the bilabial nasal pair (/m/-/mm/) received the highest discrimination accuracy (93%), followed by the voiced bilabial stop pair (/b/-/bb/; 92%). As for the Vietnamese group (Figure 5), the alveolar nasal pair (/n/-/nn/) was discriminated with the lowest discrimination accuracy (45%), while the voiced bilabial stop pair (/b/-/bb/) received the highest discrimination accuracy (62%), followed by the voiceless bilabial stop pair (/p/-/pp/; 58%).

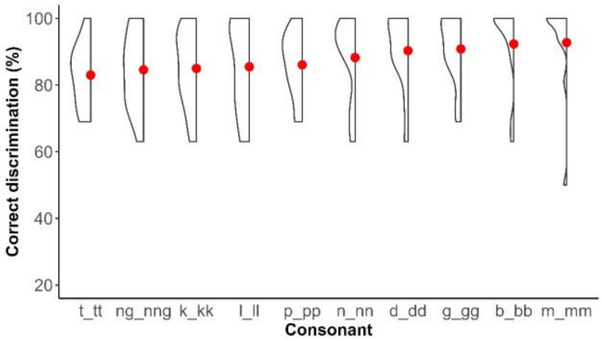


Figure 4: *Accuracy (%) of length discrimination for trials differing in the phoneme type of the target token (KM).*

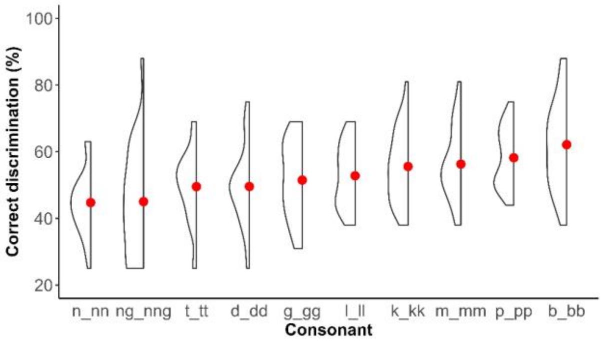


Figure 5: *Accuracy (%) of length discrimination for trials differing in the phoneme type of the target token (Vietnamese).*

4. Discussion

In this study, we examined how Vietnamese speakers may perceive the KM word-initial length contrast that is potentially challenging for non-native speakers, given the cross-linguistically marked status of this particular contrast [e.g., 5]. Consonant length is not contrastive in the Vietnamese language, though duration is employed to distinguish the tense-lax vowel distinction in this language [10]. Thus, we were interested in determining if the Vietnamese speakers, who are naïve to consonant length, are able to perceive the unfamiliar singleton/geminate contrast in word-initial position.

The findings support our earlier hypothesis: the Vietnamese speakers did face considerable difficulties in perceiving the rare word-initial consonant length contrast in KM, irrespective of the length category (singleton or geminate), with all speakers as a group discriminating the trials just above chance level (overall discrimination accuracy=53%). Surprisingly, some native speakers of KM in our study were also found to face difficulties in perceiving the length contrast in their own language, confirming the genuine difficulties of the word-initial length contrast. Our observations on Vietnamese speakers lend some evidence to the prediction by [11] who claimed that cross-language perception of phonological length contrasts would be more successful for speakers who have L1 experience with *consonant* length contrast rather than with *vowel* length contrast. Our data also accord well with the previous findings involving Vietnamese speakers such as [7], [12] and [13]. In this regard, the Vietnamese speakers in our study are comparable to American English [16] and Mandarin [15] speakers who are also less accurate in perceiving Japanese consonant length.

Pertaining to the significant effect of consonant type, the results support our earlier prediction, though different groups faced different challenges in this regard. For the KM group, the discrimination of trials beginning with voiceless stops was relatively poor, which is expected given the absence of acoustic information for listeners to discern the length contrast in this specific utterance context [e.g., 6]. With regard to the Vietnamese group, it is striking to note that trials beginning with voiceless stop pairs (e.g., /p/-/pp/) were relatively better discriminated than those beginning with other consonant types, such as nasals (see Figure 5). It can be speculated that the Vietnamese speakers in our study might have employed a unique perceptual strategy when dealing with a discrimination task involving the word-initial length contrast in KM.

5. Conclusions

The findings of the present study have shown that the speakers of Vietnamese did not match the KM native speakers in discriminating the KM singleton/geminate contrast word-initially, which is clearly due to lack of specific knowledge of the phonetic characteristics of KM singletons and geminates. Both groups were affected by the consonant type factor and discriminated KM consonant length more accurately when a specific consonant type occurred in a target position (e.g., bilabial nasals for KM participants). Our results support earlier research that experience with consonant length contrast in L1 may be helpful in processing word-initial consonant gemination.

In the future, it is obvious that there is a need to examine the perception of KM consonant length among speakers of other languages who use consonant length contrastively, such as Japanese and Italian. Further, given the complexities of word-initial consonant gemination, it would be valuable to include speakers who have experience with this particular linguistic event, such as Tashlhiyt Berber speakers, and examine if there is additional benefit of familiarity with a length contrast that occurs in word-initial position.

6. Acknowledgements

We thank Universiti Utara Malaysia (UUM) and University of Social Sciences and Humanities, Vietnam National University for their generous assistance and for making the data collection possible. We thank reviewers for their time and input.

7. References

[1] Hamzah, M. H., Fletcher, J. and Hajek, J., “Durational correlates of word-initial voiceless geminate stops: The case of Kelantan Malay”, Proceedings of the 17th International Congress of Phonetic Sciences: 815-818, 2011.

[2] Hamzah, M. H., Fletcher, J. and Hajek, J., “Word-initial voiceless stop geminates in Kelantan Malay: Acoustic evidence from amplitude/F0 ratios”, Proceedings of the 18th International Congress of Phonetic Sciences, 2015.

[3] Hamzah, M. H., Fletcher, J. and Hajek, J., “Closure duration as an acoustic correlate of the word-initial singleton/geminate consonant contrast in Kelantan Malay”, Journal of Phonetics, 58: 135-151, 2016.

[4] Hamzah, M. H., Fletcher, J. and Hajek, J., “Non-durational acoustic correlates of word-initial consonant gemination in Kelantan Malay: The potential roles of amplitude and f0”, Journal of the International Phonetic Association, 50(1): 23-60, 2020.

[5] Ridouane, R., “Geminates at the junction of phonetics and phonology”, in C. Fougeron, B. Kühnert, M. D'Imperio and N. Vallée (Eds), Laboratory Phonology 10, 61-90, De Gruyter Mouton, 2010.

[6] Ridouane, R. and Hallé, P. A., “Word-initial geminates: From production to perception”, in H. Kubozono (Ed), The Phonetics and Phonology of Geminate Consonants, 34-65, Oxford University Press, 2017.

[7] Đỗ, H. N., “Issues regarding Vietnamese learners’ perception of long vowels and geminate consonants in Japanese (in Japanese)”, Journal of Science, Foreign Languages, 31: 31-38, 2015.

[8] Kanamura, K., “Prosodic features in Japanese speech by native Vietnamese speakers (in Japanese)”, Studia Linguistica, 12: 73-91, 1999.

[9] Pajak, B. and Levy, R., “The role of abstraction in non-native speech perception”, Journal of Phonetics, 46: 147-160, 2014.

[10] Emerich, G. H., “The Vietnamese vowel system”, PhD thesis, University of Pennsylvania, US, 2012.

[11] Altmann, H., Berger, I. and Braun, B., “Asymmetries in the perception of non-native consonantal and vocalic length contrasts”, Second Language Research, 28: 387-413, 2012.

[12] Sugimoto, T., “Error analysis of Japanese pronunciation by Vietnamese learners III: On the geminate consonant and the syllabic nasal”, Departmental Bulletin of College of Humanities, Ibaraki University, 2: 149-164, 2007.

[13] Đỗ, H. N., “The influence of position and accent on the perception of long vowels and geminate consonants: Vietnamese learners majoring in Japanese (in Japanese)”, Journal of Science, Foreign Languages, 28: 242-254, 2012.

[14] Hamzah, M. H., Tsukada, K. and Hajek, J., “Perception of the Japanese word-medial singleton/geminate contrast by Kelantan Malay speakers”, in R. Skarnitzl and J. Volín (Eds), Proceedings of the 20th International Congress of Phonetic Sciences: 162-166, Guarant International, 2023.

[15] Tsukada, K., Yurong, Kim, J.-Y., Han, J.-I. and Hajek, J., “Cross-linguistic perception of the Japanese singleton/geminate contrast: Korean, Mandarin and Mongolian compared”, in H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg and P. Motlícek (Eds), Proceedings of the 22nd Annual Conference of the International Speech Communication Association (Interspeech): 3910-3914, International Speech Communication Association, 2021.

[16] Tsukada, K., Idemaru, K. and Hajek, J., “The effects of foreign language learning on the perception of Japanese consonant length contrasts”, in J. Epps, J. Wolfe, J. Smith and C. Jones (Eds), Proceedings of the 17th Australasian International Conference on Speech Science and Technology: 37-40, Australian Speech Science and Technology Association, 2018.

[17] Abramson, A. S., “The perception of word-initial consonant length: Pattani Malay”, Journal of the International Phonetic Association, 16: 8-16, 1986.

[18] Boersma, P. and Weenink, D. Praat: Doing Phonetics by Computer [version 6.4.06], retrieved from http://www.praat.org (Last viewed May 4, 2024).

[19] Strange, W. and Shafer, V. L., “Speech perception in second language learners: The re-education of selective perception”, in J. G. Hansen Edwards and M. L. Zampini (Eds), Phonology and Second Language Acquisition, 153-191, John Benjamins, 2008.

[20] R Core Team, “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria, 2019. https://www.Rproject.org/

[21] Lawrence, M. A., “ez: Easy analysis and visualization of factorial experiments”, R Package Version 4.4.0., 2016. https://CRAN.Rproject.org/package=ez

[22] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R. et al., “Welcome to the tidyverse”, Journal of Open Source Software, 4(43): 1686, 2019.

A Preliminary Study on the Vowel Length Contrast in te reo Māori

C. T. Justine Hui[1], Isabella Shields[2], Peter J. Keegan[3], Catherine I. Watson[2]

[1]Acoustics and Vibration Research Centre, University of Auckland, [2]Department of Electrical, Computer and Software Engineering, University of Auckland, [3]Te Puna Wānanga, University of Auckland

[1]justine.hui@auckland.ac.nz

Abstract

Te reo Māori (the Māori language of New Zealand) has a long-short vowel contrast. Previous studies have shown some of the long vowels have been shortened, resulting in a reduction in durational distinction between short and long vowels. Present-day elders' vowel duration was compared with younger speakers, and we found that younger speakers have shortened their short vowels, resulting in an increased vowel distinction for certain vowels. An identification perception test was carried out to examine the categorical boundary of the long-short vowel contrast for Māori learners and advanced listeners, where the advanced listeners responded more ambiguously, suggesting they may be using cues beyond duration.

Index Terms: te reo Māori, long-short vowel contrast, acoustic analysis, speech perception

1. Introduction

Vowel length distinctions are found in a number of languages such as Czech [1], Danish [2], Swedish [3], Estonian [4], Finnish [4], Thai [5], Arabic [6], Japanese [7], and Mongolian [8], where the length of the vowels can change the meaning of a word. In these languages, short vowels are phonetically realised with shorter duration, while long vowels are phonetically realised with longer duration, and the obvious primary cue to distinguish between short and long vowels is, therefore, duration in most cases. Some languages also use secondary cues such as fundamental frequency (f0) and formant values. For example, f0 is a secondary cue in vowel length perception for Japanese listeners [9–12], whereas vowel formants affect vowel length categorisation in Swedish [3]. Vowel length perception can vary across languages, where linguistic background and experience of the listener, as well as their hearing abilities, can affect how a listener perceives a vowel to be long versus short [6, 10, 12–15]. For example, second language (L2) listeners cannot identify or distinguish between vowel length contrasts when their first language (L1) does not have length contrasts [10, 13], but having length contrasts in their L1 can sometimes help in the listeners' L2 vowel length perception [6, 14].

Māori is the language of New Zealand's indigenous people, and as with all Polynesian languages, it also contrasts in vowel length. It consists of five vowels /i, e, a, o, u/, each with a distinct long vowel version where vowel length is phonemic [16–18]. Word stress will usually lie on the first long vowel if present [17, 19], resulting in long vowels being often stressed.

Throughout the past 180 years, Māori has experienced increasing contact with English, with English-medium schooling becoming prevalent by the late 19th century [17]. The subsequent decades saw a significant shift towards English dominance, resulting in a decline in fluent Māori speakers [20, 21], causing an intergenerational transmission loss of the language. Revitalisation efforts since the mid-1980s have led to a resurgence in first language (L1) speakers [22, 23]. The ongoing reversal of language shift in Māori has created a distinctive dynamic between first language (L1) and second language (L2) users [21], where many of the current generation of L1 users learn from fluent L2 speakers. The revitalisation efforts have also facilitated a growing number of L2 learners [18, 24, 25]. The Māori and New Zealand English (MAONZE) project [26, 27] examined sound changes in Māori over the past century e.g.,[16, 27–30], consisting of analyses in vowel and consonant production, as well as prosody. The MAONZE corpus consists of speakers from three generations: historical speakers born in the 1880s and recorded in the 1940s, present-day elder speakers born in the 1930s, and present-day younger speakers born around 1980, where the present-day speakers were recorded in the early 2000s [16, 26, 27].

Regarding vowel length analysis, [18] reports on the average vowel length by speaker group and notes a reduction in the duration differences between long and short vowels in Māori over time, comparing historical speakers to present-day speakers. The long vowels of the historical speakers are typically twice the length of the short vowels, but this distinction has been reduced in the present-day speakers, especially for /iː/ and /uː/, possibly as a result of New Zealand English influences. However, [18] also notes that the length distinction between /aː/ and /a/ has largely remained, possibly due to the short and long vowel pair in corresponding New Zealand English (START and STRUT vowels). To our knowledge, there has not been any further published detailed acoustic analysis of Māori vowel length apart from [18]. There has been very little perceptual research published on Māori cues [30, 31], and none on the perception of Māori vowel length.

The current paper aims to provide a more detailed analysis of Māori vowel length using the MAONZE corpus, focusing on present-day elder and younger speakers. Preliminary findings of the acoustic analysis were presented at the 2023 New Zealand Linguistics Society Conference. A preliminary perception test was also conducted to examine categorical boundaries in the duration of Māori vowel length contrasts among learners and advanced users.

2. Methodology

2.1. Corpus analysis

In the current study, only the present-day elders and first language (L1) younger speakers from the MAONZE corpus were analysed [26]. There were nine male (mean age at time of recording = 71.4, sd = 5.4) and eight female (age mean at time of recording = 73.5, sd = 7.3) speakers in the present-day elders (L1) group and five male (mean age at time of recording = 24.6, sd = 5.9) and six female (mean age at time of recording = 20.7, sd = 2.3) speakers in the younger L1 group.

The MAONZE project selected target tokens of vowels from contextual speech [27]. Thirty tokens per vowel (/i i: e e: a a: o o: u u:/) per speaker were chosen in CVC environments whenever possible. The start and end of the vowel tokens were hand labelled by the MAONZE team. The vowels selected by the MAONZE team were always stressed vowels, and they are the ones analysed in the current study. Query and duration extraction was performed using emuR [32], and linear mixed model analyses were used to analyse the duration data.

2.2. Vowel length identification

2.2.1. Stimuli

Three sets of target words were chosen using the recorded speech from the MAONZE corpus, where the speakers were instructed to say a series of hVt nonce words in the form of hV:tata and hVta. There were no minimal pair equivalents in the corpus, i.e., the short vowel equivalent to hātata was hata and not hatata. Recordings from a present-day elder female speaker affiliated with the Ngāti Pōrou iwi (tribe), born in 1930 and aged 77 at the time of recording, were selected due to the quality of the speech recorded. A present-day elder woman was chosen as they are a group reported to be more conservative in preserving vowel length distinction [33]. As young female speakers have been shown to lead sound change [16, 33–35], this was also in preparation for a future study where we will examine the differences in perception between present-day elder and young female speakers.

Table 1. *Word stimuli and their target vowel duration.*

word	target vowel duration (s)
/a:/ in ha:tata	0.17
first /a/ in hata	0.06
/e:/ in he:tata	0.20
/e/ in heta	0.09
/o:/ in ho:tata	0.17
/o/ in hota	0.07

Only three of the five vowels were chosen due to time constraints. They were: /a/ in the words hātata and hata, /e/ in the words hētata and heta and /o/ in the words hōtata and hota. The /a/ stimuli were chosen due to the reported vowel length distinction. We also included a front vowel set and a back vowel set, where hētata and heta were chosen over hītata and hita, and hōtata and hota were chosen over hūtata and huta according to the sound quality of the recordings. Table 1 shows the keywords selected and the duration of the target vowel, that is, the duration of the vowel (V) in hV:tata and hVta, measured in Praat [36] visually by the authors.

A Praat script [37] was used to automatically generate a 10-step continuum, where the short extreme value was the duration of the hVta word and the long extreme value was the duration of the hV:ta word. For example, the stimulus 'hātata' was manipulated from 0.17 in ten steps to 0.06. Similarly, 'hata' was manipulated from 0.06 in ten steps to 0.17. Due to the possibility of secondary auditory cues in perceiving long-short vowel contrast, both hV:tata and hVta continuum were included in the identification test.

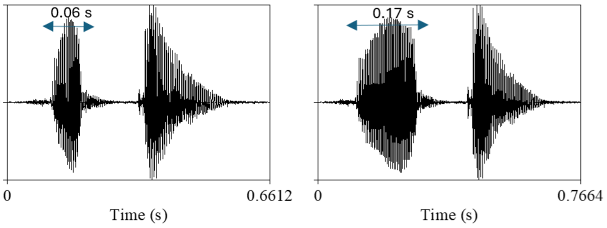


Figure 1: *Original "hata" and step-10 hata where the first /a/ has been manipulated to be as long as the /a:/ in ha:tata.*

2.2.2. Participants

Eleven te reo Māori listeners of different language experiences were recruited. They were separated into two groups: learners and advanced users. The learner's group consisted of 5 participants: 4 female listeners, 1 male listener, and no gender diverse (mean age = 27.2, sd = 6.9). The learners' group has been exposed to te reo Māori through their New Zealand schooling and/or has taken language courses in secondary/tertiary education. None speak Māori in their daily life. The advanced group consisted of 4 participants: 3 female listeners, 3 male listeners, and no gender diverse (mean age = 37.8, sd = 16.45). The advanced group includes L1 speakers, fluent L2 speakers, teachers of te reo Māori, and speakers with Māori heritage who have taken advanced tertiary courses. Recruitment and participation in perception tests have been approved by the Ethics committee at the University of Auckland (26316). They were remunerated for their participation.

2.2.3. Test design

Before the experiment started, the listeners were given practice trials to familiarise themselves with the graphical user interface (GUI) and the type of speech sounds they would be listening to. A jspsych script modified from [38] was used [39]. For each set of keywords, they listened to the two extremes of the keywords. For example, they listened to 'hata' where the /a/ vowel is as short as 0.06 s and as long as 0.17 s (refer to Table 1 and Figure 1). The participants listened to the isolated words and were asked to identify which of the two words on the screen they heard, for example, "hata" versus "hāta", "hatata" versus "hātata", where the macron above the vowel is a common practice in te reo Māori to denote a long vowel. The words were presented in blocks, with the order of the blocks randomised, and the locations of the buttons were swapped halfway through the repetitions. There were 6 repetitions per word. This gave a total of 360 trials (6 words x 10-step continuum x 6 repetitions) per listener.

3. Results

3.1. Acoustic analysis

The duration data of the labelled long and short vowels from the MAONZE corpus were statistically analysed using a linear mixed effect model via the *lme4* package in R [40]. The fixed effects were: vowel type (/i, e, a, o, u/), duration type (long vs short), gender of speaker (female vs male) and age group of speaker (elder vs younger). The fitted model includes a four-way interaction between the above fixed effects and a random effect of the speaker. Post-hoc analysis for pairwise comparison was implemented by *emmeans* package [41]. A significant four-way interaction was found between vowel type, duration type, gender of speaker and age group of speaker

Table of Contents for this manuscript

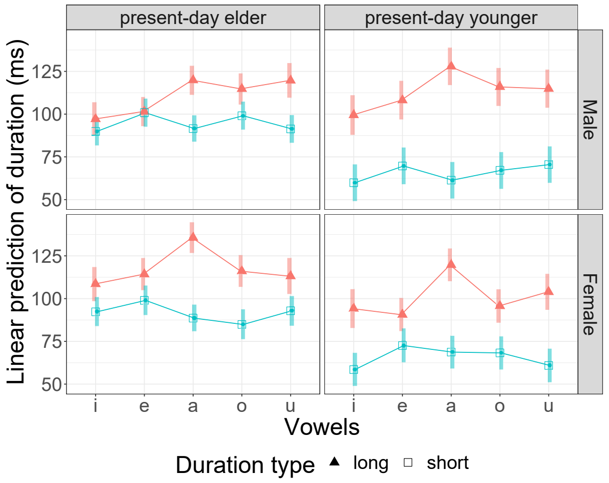


Figure 2: *Duration of vowels spoken by present-day elder and younger speakers from the MAOZNE corpus.*

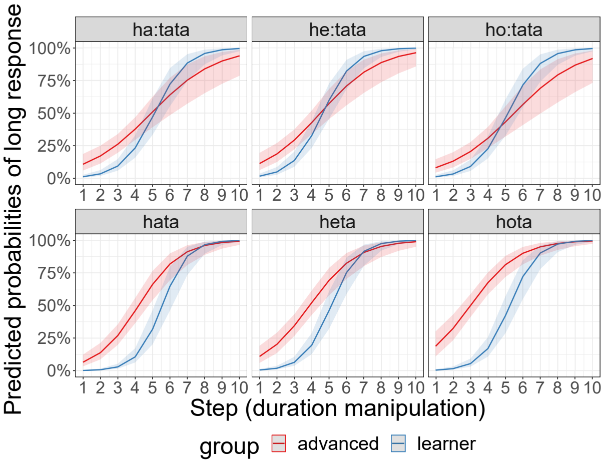


Figure 3: *Predicted long vowel responses by keyword.*

$(\chi^2(4) = 18.91, p < 0.001)$.

Figure 2 displays the duration estimates of the five long-short vowel pairs in terms of the four groups of speakers (elder male, elder female, younger male, younger female) from the fitted model. The colour and shape of the plot indicate the duration type of the vowel, where pink/triangle displays the long vowel measurements and square/blue displays the short vowel measurements. We observed that the short vowels have become shorter in the younger speakers regardless of gender. In general, differences between the short and long vowels have also become larger for the younger speakers than the elder speakers.

Due to space constraints, tables containing the pairwise contrasts from the statistical analyses are not included, and the results are described textually only. With the exception of /e/ and /i/ in the elder male speakers, all vowels produced by the four groups of speakers significantly differ in length between their short and long variants. Apart from elder male speakers, the much more frequent /a/ has the largest differences between its short and long variants compared to the other four vowels. The differences between the long and short variants of /a/ range from an estimate of 28.3 ms difference in the elder male speakers to 66.5 ms differences in the younger male speakers.

Significant differences were found between the duration of /a:/ and the other long vowels (/i:/, /e:/, /o:/, /u:/) for the two groups of female speakers. For the male speakers, there were differences between /a:/ and the front vowels (/e:/ and /i:/) for the elder male speakers, and between /a:/ and /i:/, /e:/ and /u:/ for the younger male speakers. The differences between /a:/ and the other vowels ranged from 29.11 ms between /a:/ and /e:/ for younger female speakers and 13.04 ms between /a:/ and /u:/ for the younger male speakers. For the short vowels, there were only significant differences in duration between /a/ and /e/ for the elder speakers but not for the younger speakers, where the /a/ was significantly shorter than /e/.

Between the male and female speakers compared within their respective age groups, there were significant differences between the elder male and female speakers for /a:/ and /o/, and between younger male and female speakers for /e:/ and /o:/, with the male speakers' production longer than the female speakers' production of /o/ by 14.17 ms, /e:/ by 17.52 ms and /o:/ by 20.27 ms, and shorter for /a:/ by 15.83 ms. Between the elder and younger groups comparing within the gender of the speakers, elder speakers had significantly longer short vowels than younger speakers, regardless of gender, ranging from 16.64 ms between the female elder and younger speakers' /o/ to 33.79 ms between the female elder and younger speakers' /i/. The elder female speakers also had significantly longer /a:/, /e:/ and /o:/ compared to their younger counterparts.

3.2. Identification perception results

The identification results were analysed statistically using a generalised linear mixed model (GLMM) with the R package [42] lme4 [40] using `glmer`. Two GLMM models were implemented, one by the individual keywords and another by the word form (hV:tata and hVta) and the vowel type (/a, e, o/). The other fixed effects were step (10-step continuum), group (learners vs advanced) and a by-participant random slope over each stimulus. Significance in fixed effects was determined using a likelihood ratio test by comparing a model with the effect in question with a model without the effect. Plots were created with the R package sjPlot [43] and emmeans [41].

3.2.1. By keyword

Significant two-way interactions were observed between the keyword and step $(\chi^2(5) = 24.62, p =< 0.001)$, between group and step $(\chi^2(1) = 8.28, p = 0.004)$, and between group and keyword $(\chi^2(5) = 54.51, p < 0.0001)$. The interactions can be observed in Figure 3, displaying predicted probabilities of a participant in their groups responding a long vowel to the stimulus played. The red curve represents responses from advanced listeners, while the blue curve represents responses from learners across six keywords over a ten-step continuum from short to long extremes. Advanced listeners showed less steep curves compared to learners for all keywords. Learners perceived similarly across keywords, whereas advanced listeners' responses varied by keyword. Additionally, the learners displayed a tighter confidence interval, indicating less variation within the group, whereas the advanced group exhibited a wider confidence interval, indicating more variation.

For learners, the 50% boundary point was found to be between steps 5 and 6 for all keywords except for 'he:tata'. Advanced listeners demonstrated more variations, where their responses did not form a clear sigmoid curve. The boundary for 'hata', 'heta', and 'hota' occurred earlier between steps 3 and 4 for the advanced learners. For 'ha:tata', 'he:tata', and 'ho:tata', there were no instances of 100% extreme responses (either short or long). This trend was also observed for the short extreme responses of 'hata', 'heta', and 'hota'.

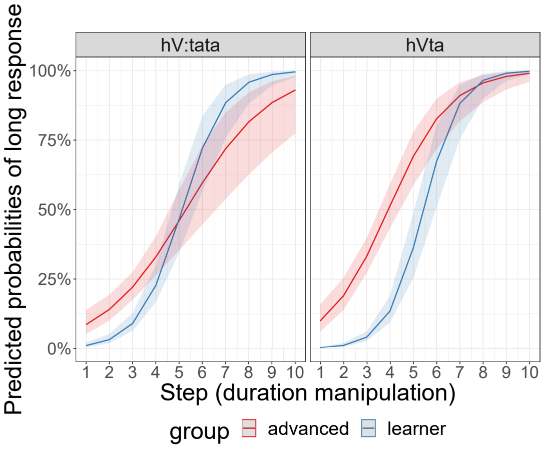


Figure 4: *Predicted long vowel responses by word form.*

3.2.2. *By word form and vowel type*

Analysis of the perceptual data by keywords indicates that advanced participants perceive hV:tata stimuli differently from hVta stimuli. This section presents an analysis based on word form and vowel rather than individual keywords. Significant two-way interactions were observed between the word form and step ($\chi^2(1) = 21.18, p < 0.0001$), group and step ($\chi^2(1) = 8.37, p = 0.004$), and group and word form ($\chi^2(1) = 48.54, p < 0.0001$), in addition to a main effect of vowel type ($\chi^2(1) = 8.72, p = 0.01$). Figure 4 displays the responses as a long vowel by the two participant groups separated by word form (hV:tata and hVta).

For the learner group, the 50% boundary is between steps 5 and 6 regardless of the word form, whereas for the advanced group, it is around step 6 for hV:tata stimuli and step 4 for hVta stimuli. This suggests that advanced listeners require more duration to perceive hV:tata as a long vowel compared to hVta, where a shorter duration elicits the perception of a long vowel.

In terms of the effect of vowels, the odds ratio contrast between /a/ and /e/ is 0.72, between /a/ and /o/ is 0.84 and between /e/ and /o/ is 1.17, regardless of the speaker groups.

4. Discussion and Conclusions

Previous analyses on vowel length contrast in Māori have noted a reduction in the durational distinction between short and long vowels in Māori. Specifically, aside from /a:/, the distinction between the vowel length pair has shortened compared to historical elders, where the duration of the long vowels was double that of the short vowels. Our analyses corroborate this trend to some extent. Compared to historical data presented in [18], present-day elder speakers exhibit shorter distinctions between the two vowel lengths. However, despite the overall reduction in durational distinction, present-day female elder speakers and male elder speakers still significantly distinguish their vowel length in terms of duration for all vowels apart from /i/ and /e/ for the male elder speakers. On the other hand, present-day younger speakers demonstrate increased durational differences between their long and short vowels compared to their present-day elder counterparts. While their long vowel durations are comparable to the elder speakers, their short vowels have reduced significantly. All short vowels among younger speakers have shortened compared to their elder counterparts, thus widening the durational distinction between the younger speakers' long and short vowel production. Changes in the present-day younger speakers may be a product of New Zealand English influences, which have been observed in other acoustic features such as vowel formants [16, 18]. Possible reasons for the reverse trend in increased durational differences observed in the younger speakers may also be hyper-articulation to compensate for the influences from New Zealand English, considering the regrets from the community regarding the intergenerational transmission loss of the language [16, 21, 25]. Future research should examine current speakers to understand how vowel lengths are realised nowadays, as the younger cohort in this study was recorded 20 years ago. Additionally, like other languages that use secondary acoustic cues such as formants and pitch contour [3, 5, 9], other acoustic cues may also be present in Māori vowel length contrast, and further acoustic analyses beyond duration is needed.

In terms of perception, we found the word and the 10-step continuum affect advanced and learner listeners differently. Word form affects advanced listeners' perception more so than learners, where the word form is how the vowel was recorded originally (i.e., the long vowels were recorded in the form of hV:tata and short vowels were recorded in the form of hVta). The advanced listeners' responses exhibit less steep slopes and more variations in general, especially for the hV:tata stimuli, indicating that the stimuli are more ambiguous to them. This ambiguity may be due to the use of nonce words as stimuli and the advanced listeners' reliance on cues other than duration. The advanced listeners may be listening for some sort of a stressed pattern when deciding whether the vowel is long or short, as word stress is often assigned to the long vowels [17]. Additionally, long vowels have also been reported to differ somewhat in vowel quality from their short variants [17], and advanced listeners may be listening for the quality as opposed to solely basing their judgement on duration, a strategy most likely to be employed by the learners. Compared to the production (acoustic analysis) part of the study, the long vowels chosen in the perception test appeared exaggerated, possibly due to the recording conditions and the use of nonce words. The /e/ stimuli elicit more long responses, likely due to the longer duration of the /e/ words selected for the study. The unusually long /e/ could be attributed to New Zealand English conflation and its status as a less frequent vowel, contributing to the speaker hyper-articulating the nonce keywords during recording.

The speaker of the stimuli was a present-day elder speaker, who, as shown in the acoustic analyses part of the study had a smaller distinction between long and short vowel contrasts compared to present-day younger speakers for certain vowels. Another possibility for advanced listeners to have more ambiguous responses is that they may be employing different listening strategies based on the age and type of speaker as they have more experience listening to Māori. One major limitation of the current study is the small number of listeners in the advanced group, and within the group, there was a wide range of experiences and daily Māori usage.

These observations suggest that advanced listeners are utilising cues that learners may not be using. Further acoustic feature analysis and production studies are needed to understand these differences. A cue-weighting approach to manipulate duration and other acoustic features such as the fundamental frequency in the stimuli could provide deeper insights into the perceptual strategies used by different listeners. As we had a diverse group of advanced listeners, more L1 and fluent L2 listeners are needed to understand the effect of language experiences on how Māori vowel length is perceived.

5. Acknowledgements

This research has been supported by the Marsden fast-start grant, managed by Royal Society Te Apārangi (22-UOA-172). We also give our sincere thanks to our wonderful participants.

6. References

[1] V. J. Podlipský, K. Chládková, and Š. Šimáčková, "Spectrum as a perceptual cue to vowel length in Czech, a quantity language," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, EL352–EL357, 2019, ISSN: 0001-4966.

[2] D. J. Morris and H. Juul, "The long and the short of vowel length perception in Danish," *The Journal of the Acoustical Society of America*, vol. 152, no. 5, pp. 2953–2961, 2022.

[3] D. Behne, T. A. Arai, B. Czigler, and K. Sullivan, "Vowel Duration and Spectra As Perceptual Cues To Vowel Quantity: a Comparison of Japanese and Swedish," *International Congress of Phonetic Sciences (ICPhS)*, pp. 857–860, 1999.

[4] I. Lehiste, "The Function of Quantity in Finnish and Estonian," *Language*, vol. 41, no. 3, pp. 447–456, 1965.

[5] A. S. Abramson and N. Reo, "Distinctive vowel length: duration vs. spectrum in Thai," *Journal of Phonetics*, vol. 18, no. 2, pp. 79–92, 1990, ISSN: 00954470.

[6] K. Tsukada, "The perception of Arabic and Japanese short and long vowels by native speakers of Arabic, Japanese, and Persian.," *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 989–998, 2011, ISSN: 00014966.

[7] L. Labrune, *The phonology of Japanese*. Oxford: Oxford University Press, 2012.

[8] A. Min, D. Baiyila, and A. Li, "Production and perception of long and short vowel contrast in Mongolian," in *International Congress of Phonetic Science (ICPhs2023)*, 2023, pp. 302–306.

[9] K. Kinoshita, D. M. Behne, and T. Arai, "Duration and F0 as perceptual cues to Japanese vowel quantity," *Proceedings of the 7th International Conference on Spoken Language Processing*, no. 1, pp. 757–760, 2002.

[10] H. Lehnert-LeHouillier, "A cross-linguistic investigation of cues to vowel length perception," *Journal of Phonetics*, vol. 38, no. 3, pp. 472–482, 2010, ISSN: 00954470.

[11] I. Takiguchi, H. Takeyasu, and M. Giriko, "Effects of a dynamic F0 on the perceived vowel duration in Japanese," *Speech Prosody*, pp. 14–17, 2010, ISSN: 23332042.

[12] C. T. J. Hui and T. Arai, "Elderly listeners' identification of Japanese long vowel pair 'obasan' and 'obaasan' using pitch and duration," *Acoustical Science and Technology*, vol. 40, no. 2, pp. 105–115, 2019, ISSN: 1346-3969.

[13] S. Ylinen, A. Shestakova, P. Alku, and M. Huotilainen, "The perception of phonological quantity based on durational cues by native speakers, second-language users and nonspeakers of finnish," *Language and Speech*, vol. 48, no. 3, pp. 313–338, 2005, ISSN: 00238309.

[14] I. Takiguchi, "The role of vowel duration cue in L1: Effects on L2 learners' identification of phonological vowel length in Japanese," *ICPhS*, pp. 1–5, 2015.

[15] C. T. J. Hui and T. Arai, "Pitch and duration as auditory cues to identify Japanese long vowels for Japanese learners," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 797–799, 2020.

[16] C. I. Watson, M. A. Maclagan, J. King, R. Harlow, and P. J. Keegan, "Sound change in Māori and the influence of New Zealand English," *Journal of the International Phonetic Association*, vol. 46, no. 2, pp. 185–218, 2016, ISSN: 14753502.

[17] R. Harlow, *Māori - A linguistic introduction*. Cambridge: Cambridge University Press, 2007.

[18] J. King, R. Harlow, C. I. Watson, P. J. Keegan, and M. Maclagan, "Changing pronunciation of the Māori language implications for revitalization," *Indigenous Language Revitalization: Encouragement, Guidance and Lessons Learned*, pp. 85–96, 2009.

[19] B. Biggs, *Let's Learn Māori : A Guide to the Study of the Māori Language*. Reed, 1969.

[20] R. A. Benton, "The Maori Language: Dying or reviving?" New Zealand Council for Educational Research., Wellington, Tech. Rep., 1997, p. 47.

[21] J. A. Fishman, "Maori: The Native Language of New Zealand Background," in *Reversing Language Shift: Theoreticdal and Empirical Foundations of Assistance to Threatened Languages*, 1991, p. 19.

[22] R. Benton and N. Benton, "RLS in Aotearoa / New Zealand 1989 – 1999," in *Can Threatened Languages Be Saved?* J. A. Fishman, Ed., Multilingual Matters, 2001, ch. 18, pp. 423–450.

[23] R. Harlow and J. Barbour, "Māori in the 21st Century: climate change for a minority language?" In *Language ecology for the 21st century: Linguistic conflicts and social environments*, 2013, pp. 241–266.

[24] S. May, "Introduction bilingual/immersion education in Aotearoa/New Zealand: Setting the context," *International Journal of Bilingual Education and Bilingualism*, vol. 8, no. 5, pp. 365–376, 2005, ISSN: 13670050.

[25] P. J. Keegan, C. I. Watson, J. King, M. Maclagan, and R. Harlow, "The role of technology in measuring changes in the pronunciation of Māori over generations," in *Language Endangerment in the 21st Century: Globalisation, Technology and New Media, FEL XVI*, Auckland, 2012.

[26] J. King, M. Maclagan, R. Harlow, P. Keegan, and C. Watson, "The Maonze Corpus: Establishing A Corpus Of Maori Speech," *New Zealand Studies in Applied Linguistics*, vol. 16, no. 2, pp. 1–16, 2010, ISSN: 1173-5562, 1173-5562.

[27] J. King, M. Maclagan, R. Harlow, P. J. Keegan, and C. I. Watson, "The MAONZE Corpus: Transcribing and analysing Maori speech," *New Zealand Studies in Applied Linguistics*, vol. 17, no. 1, pp. 32–48, 2011, ISSN: 1173-5562, 1173-5562.

[28] P. J. Keegan, J. King, M. Maclagan, C. I. Watson, and R. Harlow, "Changes in the pronunciation of Māori and implications for teachers and learners of Māori," in *LED*, 2009.

[29] M. Maclagan, R. Harlow, J. King, P. J. Keegan, and C. I. Watson, "Acoustic Analysis of Māori: Historical Data," *Proceedings of the 2004 Conference of the Australian Linguistic Society*, no. October, pp. 1–16, 2004.

[30] L. Thompson *et al.*, "Adventures in Mita-Reading : Examining Stress ' Rules ' and Perception of Prosodic Prominence in the Maori Language," *ICPhS*, no. February 2016, pp. 1–4, 2011.

[31] C. I. Watson *et al.*, "Prosodic clues in language recognition: how much information do listeners need to identify Maori and English?" *Te Reo*, vol. 54, p. 83, 2011.

[32] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech and Language*, vol. 45, pp. 392–410, 2017, ISSN: 10958363.

[33] J. King, M. Maclagan, R. Harlow, P. Keegan, and C. Watson, "Prestige norms and sound change in Māori," *Language Ecology*, vol. 4, no. 1, pp. 95–114, 2020, ISSN: 2452-1949.

[34] J. Holmes, "Maori and Pakeha English : Some New Zealand Social Dialect," *Language in Society*, vol. 26, no. 1, pp. 65–101, 1997.

[35] J. Holmes, "Setting New Standards: Sound Changes and Gender in New Zealand English," *English World-Wide*, vol. 18, no. 1, pp. 107–142, 1997.

[36] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, 2016. [Online]. Available: `http://www.fon.hum.uva.nl/praat/`.

[37] M. B. Winn, *Make duration contiuum*, 2014. [Online]. Available: `http://mattwinn.com/praat.html#durationContinuum`.

[38] L. Sullivan, *Basic forced choice task*, 2020. [Online]. Available: `https://lisasullivan.ca/sample_experiments/forced_choice_no_feedback.html`.

[39] J. R. de Leeuw, R. A. Gilbert, and B. Luchterhandt, "Jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments," *Journal of Open Source Software*, vol. 8, p. 5351, 85 May 2023.

[40] D. Bates, *Linear mixed model implementation in lme4*, 2007.

[41] R. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2019. [Online]. Available: `https://cran.r-project.org/package=emmeans`.

[42] R Core Team, *R: A language and environment for statistical computer*, 2015. [Online]. Available: `https://www.r-project.org/`.

[43] D. Lüdecke, *sjPlot: Data visualization for statistics in social science*, Available at https://cran.r-project.org/package=sjPlot, 2018. [Online]. Available: `https://cran.r-project.org/package=sjPlot`.

Vowel Duration beyond Contrastive Length in Djambarrpuyŋu

Kathleen Jepson[1,2], *Rasmus Puggaard-Rode*[2]

[1]University of Queensland; [2]Institute of Phonetics and Speech Processing, LMU Munich

k.jepson@uq.edu.au; r.puggaard@phonetik.uni-muenchen.de

Abstract

Djambarrpuyŋu has contrastive vowel length only in word-initial syllables; long vowels are twice as long as short vowels. Little is known, however, about vowel duration outside of this position. Vowel duration can be affected by various other factors including syllable structure, number of syllables in the word, and proximity to prosodic boundaries, topics that require further investigation in Djambarrpuyŋu. This paper aims to describe the durational patterns of vowels, beyond contrastive length, to enhance our understanding of vowel duration and factors that affect it in Djambarrpuyŋu, and to contribute to our cross-linguistic understanding of segment duration.

Index Terms: vowels, duration, syllable structure, polysyllabic shortening, final lengthening, vowel length, Djambarrpuyŋu

1. Introduction

Djambarrpuyŋu has contrastive vowel length, but only in word-initial syllables [1]. This is the proposed location of primary stress, which is fixed (see Section 1.1). Long vowels are, on average, twice as long as short vowels with duration values of ~200 ms and ~100 ms respectively [2]. However, little is known about the durational characteristics beyond these word-initial, stressed vowels, and a more nuanced understanding of the effect of other factors on vowel duration awaits further investigation. Non-contrastive vowels, which have not been investigated acoustically, are typically transcribed using a plain vowel symbol, categorising them in this way as short vowels, although it is not known if they actually pattern with short vowels in terms of their durational characteristics.

Across languages, it is understood that vowel duration can be affected by a number of factors such as syllable structure [3], number of syllables in the word [4, 5], and proximity to prosodic domain boundaries [6, 7]. The aim of this paper is to deepen our understanding of Djambarrpuyŋu vowels, including non-contrastive vowels, and in doing so to contribute to what is known about the durational characteristics of non-contrastive vowels in a language with a length contrast, a topic which is not well-understood. Specifically, we investigate the effects of syllable structure on vowel duration, effects of polysyllabic shortening across contrastive and non-contrastive vowels, and durational modification due to proximity to the word-final boundary.

The cross-linguistic literature and expectations for the effects of these factors in Djambarrpuyŋu are discussed in Sections 1.2–1.4. Djambarrpuyngu is introduced further in Section 1.1.

1.1. The Djambarrpuyŋu language

Djambarrpuyŋu, a Pama-Nyungan language spoken by ~4,000 people in northeast Arnhem Land [8], is described as having six contrastive vowels /ɪ ɪː ɐ ɐː ʊ ʊː/ [1]. While it is simplest to describe contrastive length as being restricted to word-initial (i.e., stressed) syllables, long vowels can occur in word-medial position in compounds when the second compound member has a long vowel in the initial syllable, for example, *yaŋara'-märrma'* /ˈjɐ.ŋɐ.ɹɐʔ.mɐːr.mɐʔ/ “twins” (lit. “lower leg two”) [1]. However, long vowels are proposed to be shortened in the second morpheme of reduplicated stems, for example, *yolŋu'-yulŋu* /ˈjʊːl.ŋʊʔ.jʊl.ŋʊ/ “people”. In the current analysis, vowels at the beginning of the second morpheme in compounds are coded as being either short or long, and these are considered “medial” with respect to their position within the word. Vowels in the word-initial syllable are coded as either short or long, vowels elsewhere within the word are coded as non-contrastive.

Like many Australian languages, Djambarrpuyŋu has relatively free word order, and is a highly agglutinating, exclusively suffixing language [1]. Syllables take the form CV(C)(C)(ʔ), where C represents a consonant and V represents a vowel. Syllable onsets are always a single consonant, codas can be more complex, glottal stops only occur syllable-finally.

Acoustic and perceptual investigations have examined vowel length in disyllabic words, including the effect of syllable structure [2]. The effect of final lengthening on consonants has also been considered [9]. Details of the findings are discussed in the following sections.

1.2. Effect of syllable structure

A common pattern observed cross-linguistically is for vowels in closed syllables such as CVC to be phonetically shorter than in open syllables such as CV [3]. This effect has been reported to occur in languages that have contrastive vowel length such as Dutch [10, 11], Arabic [12-14], and Malayalam [12], as well as those which do not such as Italian [15, 16]. In Dutch for example, short vowels are 82 ms in open syllables and 51 ms in closed syllables, and long vowels are 178 ms in open syllables and 124 ms in closed syllables [11].

The effect of syllable structure on vowel duration in disyllabic Djambarrpuyŋu words is reported in [2]. Closed syllables did affect the duration of vowels; however, long vowels were affected to a greater extent than short vowels. Long vowels in closed syllables are approximately three quarters the duration of long vowels in open syllables. Therefore, the ratio between short and long vowels is altered from ~1:2 in open syllables to 1:1.5 in closed syllables.

It is not yet known what effect syllable structure has on non-contrastive vowels in Djambarrpuyŋu, though it is expected that non-contrastive vowels behave in a similar way to short vowels, and that closed syllable vowel shortening results in only slightly shorter non-contrastive vowels.

1.3. Effect of polysyllabic shortening

Polysyllabic shortening is a mechanism whereby syllable duration, especially that of primary stressed syllables in accented words, is negatively correlated with the number of syllables in the word [4, 5, 17-19]. [4] examined the effect of polysyllabic shortening in English. They observed polysyllabic shortening is stronger for accented words than unaccented words irrespective of location of main stress. In German it has similarly been found that stressed vowels in accented words show an effect of polysyllabic shortening [5]. The effect was observed for tense vowels but not lax vowels, resulting in a smaller difference in duration between the two categories. Polysyllabic shortening is reported to not occur in some languages, however. For example, the number of syllables in a word has no consistent effect on segment duration in Finnish [20].

Based on the cross-linguistic literature, polysyllabic shortening is expected to be observed in Djambarrpuyŋu across vowel categories and positions. However, long vowels are expected to be affected to a greater degree than short vowels in stressed position, and short vowels would in turn be affected to a greater extent than non-contrastive vowels, which do not occur in stressed position. It is expected that long vowels in open and closed syllables could be affected differently by polysyllabic shortening. Using the current data set it is not possible to compare accented versus unaccented words, so additional effects due to accentuation awaits further investigation.

1.4. Effect of final lengthening

Domain-final lengthening is the phonetic lengthening of segments due to proximity to a prosodic constituent boundary, and is often found to affect syllable rimes, that is, syllable-final vowel-consonant sequences [6, 7].

For English, [21] report that final lengthening affects word-final rimes, and also main stress syllable rimes of phrase-final words. That is, domain-final lengthening was found to affect non-final stressed segments in domain-final words. In Dutch, word-final consonants as well as the preceding vowel are found to be affected by being in utterance-final position, with consonants, not vowels, contributing the majority of the durational difference between syllables across positions [22].

Most research has focused on higher-level prosodic constituent boundaries such as the Intonational Phrase (IP), and the varying degrees of lengthening corresponding to different levels of the prosodic hierarchy; lower level boundaries (e.g., Prosodic Word) result in a small degree of lengthening, with increases in lengthening for higher-level constituent boundaries (e.g., IP, Utterance) [21].

IP-final lengthening of vowels is anecdotally reported to occur in Djambarrpuyŋu [23]. Further, consonants have been found to be lengthened when adjacent to a prosodic phrase boundary in Djambarrpuyŋu, with nasals followed by a break being ~ 55 ms longer than when not [9]. However, in this paper, we are concerned primarily with the domain of the word and focus on the effect of lengthening in final versus non-final syllables. Considering only non-contrastive vowels, we hypothesise that vowels in word-final syllables are longer than vowels elsewhere within the word, and that syllable structure minimally affects the effect of final lengthening.

2. Methods

2.1. Participants

Eight Djambarrpuyŋu speakers (five women, three men) were recorded in Milingimbi, northeast Arnhem Land, Northern Territory, Australia. All participants were familiar with related language varieties, other Aboriginal languages, and Australian English. Participants were paid for their time.

2.2. Materials and recordings

A wordlist was compiled making use of grammatical [1] and dictionary resources [24]. Words included all vowels, had varying morphological structure including compounds and reduplicated forms, and were between one and eight syllables in length. A total of 8,995 vowels are examined in the analysis, the distribution of these is presented in Table 1. Note that vowels from monosyllabic words are included in Table 1 and Figures for interest, but are not included in the statistical analyses.

The wordlist items were elicited in three frame sentences in which the target word was syntactically in utterance-initial, -medial, or -final position. The items in the wordlist were discussed with each speaker before the recording session. In the recording session, each item was presented verbally in English, Djambarrpuyŋu, or through an explanation in English. Speakers said each item in the three frame sentences once. Sentence frame is not examined in detail in this paper, nor was the occurrence of a pause following the target word (i.e., a proxy for a prosodic constituent boundary), though it is acknowledged that proximity to higher-level prosodic boundaries would affect vowel duration, and this will be explored imminently.

Audio data were collected using a Zoom H6 digital recorder and Countryman H6 headset microphone with a hypercardioid pattern directional capsule covered with a windshield. Recordings were made at 24 bit bit-depth and a 48 kHz sample rate. Recording sessions primarily took place sitting inside a house, with over-head fans and air-conditioning units turned off, or outside in the shade or on a veranda.

Table 1. *Summary of the data presented by vowel length category, position within the word, and structure of the syllable counting vowel tokens, and by number of syllables in the word counting word tokens.*

Vowel	*V n*	Pos. in word	*V n*
long	732	*initial*	2838
short	2237	*medial*	3294
non-contr.	6026	*final*	2838
		mono	25
Syll. structure	*V n*	Sylls. in word	*W n*
open	6380	*1*	25
closed	2615	*2*	1554
		3	383
		4	214
		5	355
		6	255
		7	53
		8	24

2.3. Data processing and acoustic measures

Utterance segmentation and transcription were conducted in Praat [25] using a modified Djambarrpuyŋu orthography. Data were forced aligned on two separate occasions and two slightly differing methods were used. For the first set of data, the Munich Automatic Segmentation System [26] implemented in R [27] was employed, using a modified SAMPA (language-independent) parameter definition. For the second set of data, the web-based Munich Automatic Segmentation System [28] was employed, using the language-independent model. All segmentation was manually corrected in Praat using waveform and spectral information.

An Emu SDMS database was created using the emuR package in R [29]. The database was queried in R using the emuR suite of commands. Durational values were extracted for all vowels along with other relevant information.

2.4. Analysis

Linear mixed-effects regression models were fitted to test the predictions made in Sections 1.2–1.4, using the lme4 library [30] in R. Two separate models were fitted. Visualisations were created using the ggplot2 package in R [31].

Since syllable structure and polysyllabic shortening potentially interact, we fitted a model with vowel duration as the dependent variable, predicted from a three-way interaction between syllable structure (*open* or *closed*), phonological vowel length (*long*, *short*, or *non-contrastive*), and number of syllables (numeric, between 2–8). The model includes uncorrelated random by-speaker slopes for syllable structure and phonological vowel length, and random by-item intercepts; models with more elaborate random effects structures failed to converge, and the random effects structure was pruned to remove by-speaker slopes with very low predicted variance. The data for this model does not include vowels in final syllables ($n = 6132$).

A separate model tested the effects of final lengthening. This model also had vowel duration as the dependent variable, predicted from a two-way interaction between syllable finality (a binary variable), and syllable structure. The model included random by-speaker slopes for both syllable finality and syllable structure (but not their interaction), and random by-item intercepts. The random effects structure was pruned as in the previous model. The data for this model only includes non-contrastive vowels, as vowels with contrastive length never appear word-finally ($n = 6026$).

Individual comparisons of interacting variables are obtained using the emmeans [32] library in R, which reports Bonferroni-corrected p-values. Due to the large number of comparisons, only selected ones are reported here. p-values < 0.05 are considered statistically significant.

3. Results

Figures 1 and 2 present the data in different organisations that relate to the statistical analyses, although the figures present all data, including that which was excluded from the statistical models. Fig. 1 shows vowel duration (ms) by position within the word (*initial*, *medial*, *final*, *monosyllabic*) and phonological length category of the vowel (*short*, *long*, *non-contrastive*), coloured by syllable structure (*open*, *closed*). Vowels in monosyllabic words could arguably be considered either initial or final vowels, and so are their own category in the plots for descriptive information for the reader. Fig. 2 shows the number of syllables per word (1-8) by phonological length category, coloured by syllable structure.

3.1. Syllable structure

The effect of syllable structure can be observed in both Figures 1 and 2. As has previously been reported [2, 33], we found that long vowels are shorter in closed syllables (yellow) than open syllables (blue), while the duration of short and non-contrastive vowels is not affected to a considerable degree by syllable structure. The effect for the long vowel data is observed in the location of the peaks in the distribution in the facet representing initial long vowels in Fig. 1 and long vowels in disyllabic words in Fig. 2. The statistical analysis finds that open long vowels are an estimated 17 ms (95% CI = [7, 27]) longer than closed long vowels; this difference is significant with $p < 0.001$. Syllable structure does not significantly affect the duration of short and non-contrastive vowels. The effect of syllable structure on vowels in word-final syllables is discussed in Section 3.3.

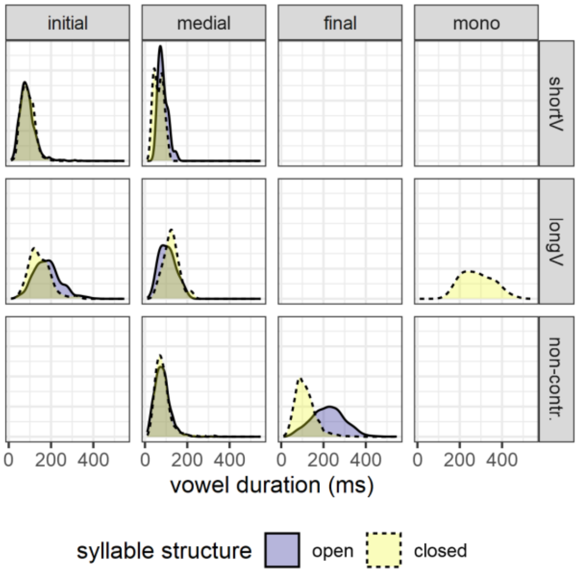


Figure 1: *Duration of vowels by position within the word and length category, coloured by whether the syllable was open or closed.*

3.2. Polysyllabic shortening

Fig. 2 shows the effect of number of syllables in the word on vowel duration. Considering short vowels first, there is little change in duration across words of different lengths in open or closed syllables; they are usually in the range of 100 ms. Next, considering long vowels, there is less data, but overall what can be observed is that vowel durations in open syllables (blue) become shorter, and more similar to vowels in closed syllables (yellow). Lastly, non-contrastive vowels in 3–8 syllable words show the same pattern as short vowels. That is, the duration values across word lengths and syllable structures are nearly identical, and hover around 100 ms. Non-contrastive vowels in disyllabic words are considered more in Section 3.3; these vowels are necessarily in word-final syllables.

The statistical analysis predicts that with the addition of each subsequent syllable, short vowels shorten by -5.6 ms, 95% CI = [-3.8, -7.3]. The effect of increasing syllable count does not significantly differ between short and non-contrastive vowels, nor does syllable structure significantly affect polysyllabic shortening in short and non-contrastive vowels. In long vowels in open syllables, there is a stronger effect of polysyllabic shortening, with a predicted decrease in duration for each added syllable of -9.3 ms, 95% CI = [-2.3, -16.3], and the effect is much stronger in long vowels in closed syllables, with a predicted decrease of -24.3 ms (95% CI = [-21.1, -27.5]) for each added syllable.

3.3. Final lengthening

In this study, all vowels that are word-final are non-contrastive, so, as mentioned above, the analysis only considers non-contrastive vowels across word positions.

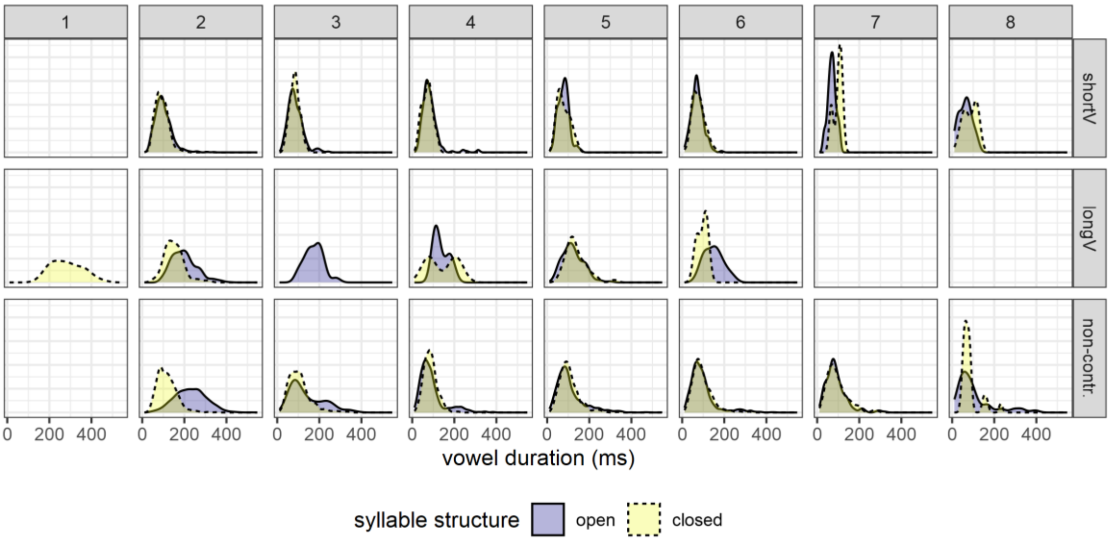


Figure 2: *Duration of vowels by the number of syllables in the word (1-8) and length category, coloured by whether the syllable was open or closed.*

The bottom row of Fig. 1 represents the data in this analysis. As can be seen, for medial (i.e., the non-final category) vowels, duration values for vowels in closed and open syllables are nearly identical; this was supported in the analysis of syllable structure, presented in Section 3.1. In word-final position, only vowels in open syllables appear to show a lengthening effect, contrary to the hypothesis that syllable structure would have little effect on vowel duration. The figure clearly shows that vowels in open syllables (blue) are longer than vowels in closed syllables (yellow) in final position.

The statistical analysis supports this, predicting that vowels in word-final open syllables are significantly longer than all other vowels by over 100 ms, $p < 0.001$.

4. Discussion and conclusions

This paper provides an overview of vowel duration in Djambarrpuyŋu and the effects of factors that are frequently found to affect vowel duration. Contrastive vowel length was considered, but oft neglected non-contrastive vowels were also examined with respect to the effect of syllable structure, polysyllabicity, and final lengthening.

For syllable structure, previous findings were replicated across words of varying lengths. Long vowels are affected by syllable structure to a greater degree than short vowels, with vowels in closed syllables shorter than those in open syllables. The hypothesis that non-contrastive vowels pattern with short vowels and show minimal effect of syllable structure was generally supported. Data for long vowels in longer words is sparse, but overall, we observe the expected effect of syllable structure across words of different lengths. A next step could be investigating the effect of different codas to understand if more complex codas result in different degrees of shortening. For now, no differentiation was made between codas in terms of how many consonants they contained, nor consonant manner.

Long vowels were affected by polysyllabic shortening to a greater degree than short vowels and non-contrastive vowels. Further, long vowels in open syllables were affected to the greatest degree with each additional syllable. On the whole, short vowels (in stressed position) did not appear to be affected to a greater degree than non-contrastive vowels. This suggests that polysyllabic shortening selectively applies to stressed vowels, as also found in German [5].

Regarding final lengthening, it was hypothesised that word-final vowels would be longer than non-contrastive vowels elsewhere in the word, and that syllable structure would not affect the duration. Word-final position did result in longer duration values. However, this was due wholly to vowels in open syllables; vowels in closed syllables were only minimally lengthened in final position. Taken together with findings from [9], this suggests that final lengthening may only be observed on absolute final segments. Exploring the effect of different prosodic constituent boundaries on vowel duration remains for future research.

Some other interesting findings emerged that were not explicitly examined. In this dataset, monosyllabic words always contain long vowels and are closed syllables. These vowels are the longest in the dataset overall. This is at once expected, but also surprising in light of other findings. We might expect long duration values because 1) these are long vowels, 2) there can be no effect of polysyllabic shortening. Further, word-final position resulted in significantly longer duration for non-contrastive vowels. However, being in a closed syllable resulted in a near total reduction of the effect of final position for non-contrastive vowels. Inclusion of monosyllabic words without codas is required to better understand these durational patterns.

Incidentally, we can also see in Fig. 1 that medial long vowels (i.e., long vowels in the second member of a compound), are considerably shorter than initial long vowels. This presents an opportunity for further exploration of the duration of these vowels, and reconsidering the phonological representation of vowels in compounds.

Overall, it appears that durationally, non-contrastive vowels pattern with short vowels, with similar effects of syllable structure and polysyllabicity. Non-contrastive vowels in word-final position, however, are in the range of duration values observed for long vowels. Syllable structure affects both long vowels and word-final non-contrastive vowels in a similar way, but long vowels in closed monosyllabic words remain long in duration. This study illustrates how multiple factors that affect vowel duration can operate simultaneously and differently on vowels, and disentangling these factors is important in fully understanding a language's durational patterns.

Table of Contents for this manuscript

5. References

[1] Wilkinson, M., *Djambarrpuyŋu: A Yolŋu variety of northern Australia*. Munich: LINCOM Europa, 2012.

[2] Jepson, K., "Contrastive vowel length and segment duration in production and perception in Djambarrpuyŋu", in revision.

[3] Maddieson, I., "Phonetic cues to syllabification", in V. A. Fromkin [Ed.], *Phonetic linguistics: Essays in honor of Peter Ladefoged*, Orlando: Academic Press, 1985, 203-221.

[4] White, L. and Turk, A., "English words on the Procrustean bed: Polysyllabic shortening reconsidered", *Journal of Phonetics,* vol. 38, no. 3, 459-471, 2010, doi: 10.1016/j.wocn.2010.05.002.

[5] Siddins, J., Harrington, J., Kleber, F., and Reubold, U., "The influence of accentuation and polysyllabicity on compensatory shortening in German", presented at INTERSPEECH 2013, Lyon, France, 25-29 August, 2013.

[6] Fletcher, J., "The prosody of speech: Timing and rhythm", in W. J. Hardcastle, J. Laver, and F. E. Gibbon [Eds.], *The handbook of phonetic sciences*, Chichester: Blackwell Publishers, 2010, 523-602.

[7] Cho, T., "Prosodic boundary strengthening in the phonetics-prosody interface", *Language and Linguistics Compass,* vol. 10, no. 3, 120-141, 2016, doi: 10.1111/lnc3.12178.

[8] Australian Bureau of Statistics. "Cultural diversity: Census." https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#data-download, accessed 13 March 2023.

[9] Jepson, K., Fletcher, J., and Stoakes, H., "Prosodically conditioned consonant duration in Djambarrpuyŋu", *Language and Speech,* Special Issue: Prosodic prominence – a cross-linguistic perspective, vol. 64, no. 2, 261-290, 2021, doi: 10.1177/0023830919826607.

[10] Rietveld, T. and Frauenfelder, U. H., "The effects of syllable structure on vowel duration", in *Proceedings of the 11th International Congress of Phonetic Sciences*, Tallinn, 1987, 28-31.

[11] Jongman, A., "Effects of vowel length and syllable structure on segment duration in Dutch", *Journal of Phonetics,* vol. 26, 207-222, 1998, doi: 10.1006/jpho.1998.0075.

[12] Broselow, E., Chen, S.-I., and Huffman, M. K., "Syllable weight: Convergence of phonology and phonetics", *Phonology,* vol. 14, no. 1, 47-82, 1997, doi: 10.1017/S095267579700331X.

[13] de Jong, K. and Zawaydeh, B., "Stress, duration, and intonation in Arabic word-level prosody", *Journal of Phonetics,* vol. 27, no. 1, 3-22, 1999, doi: 10.1006/jpho.1998.0088.

[14] Khattab, G., "A phonetic study of gemination in Lebanese Arabic", in Germany, J. Trouvain and W. J. Barry, [Eds.], *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 2007, 153-158.

[15] Farnetani, E. and Kori, S., "Effects of syllable and word structure on segmental durations in spoken Italian", *Speech Communication,* vol. 5, no. 1, 17-34, 1986, doi: 10.1016/0167-6393(86)90027-0.

[16] Hajek, J., Stevens, M., and Webster, G., "Vowel duration, compression and lengthening in stressed syllables in Italian", in *Proceedings of the International Congress of Phonetic Sciences XVI,* Saarbrücken, 2007, 1057-1060.

[17] Crystal, T. H. and House, A. S., "Articulation rate and the duration of syllables and stress groups in connected speech", *The Journal of the Acoustical Society of America,* vol. 88, 101-112, 1990, doi: 10.1121/1.399955.

[18] Turk, A., "The temporal implementation of prosodic structure", in A. C. Cohn, C. Fougeron, M. K. Huffman, and M. E. L. Renwick [Eds.], *The Oxford handbook of laboratory phonology*, Oxford: Oxford University Press, 2012, 242-253.

[19] Turk, A. and Shattuck-Hufnagel, S., "Word-boundary-related duration patterns in English", *Journal of Phonetics,* vol. 28, no. 4, 397-440, 2000, doi: 10.006/jpho.2000.0123.

[20] Suomi, K., "On the tonal and temporal domains of accent in Finnish", *Journal of Phonetics,* vol. 35, no. 1, 40-55, 2007, doi: 10.1016/j.wocn.2005.12.001.

[21] Turk, A. and Shattuck-Hufnagel, S., "Multiple targets of phrase-final lengthening in American English words", *Journal of Phonetics,* vol. 35, no. 4, 445-472, 2007, doi: 10.1016/j.wocn.2006.12.001.

[22] Hofhuis, E., Gussenhoven, C., and Rietveld, T., "Final lengthening at prosodic boundaries in Dutch", in *Proceedings of the International Congress of Phonetic Sciences*, Stockholm, 1995, vol. 1, 154-157.

[23] Jepson, K. and Fletcher, J., "The intonation of Djambarrpuyŋu" in S.-A. Jun and S. D. Khan [Eds.], *Prosodic Typology III*, Oxford: Oxford University Press, accepted.

[24] Greatorex, J. Yolŋu Matha Dictionary [Online] Available: http://yolngudictionary.cdu.edu.au/

[25] *Praat: Doing phonetics by computer*. (2023). Version 6.4.01. [Online]. https://fon.hum.uva.nl/praat

[26] Schiel, F., Draxler, C., and Harrington, J., "Phonemic segmentation and labelling using the MAUS technique", presented at the New Tools and Methods for Very-Large-Scale Phonetics Research, University of Pennsylvania, 28-31 January, 2011.

[27] *R: A language and environment for statistical computing*. (2023). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org/

[28] Kisler, T., Reichel, U. D., and Schiel, F., "Multilingual processing of speech via web services", *Computer Speech & Language,* vol. 45, 326-347, 2017, doi: 10.1016/j.csl.2017.01.005.

[29] *emuR: Main Package of the EMU Speech Database Management System*. (2023). [Online]. Available: https://CRAN.R-project.org/package=emuR

[30] Bates, D., Mächler, M., Bolker, B., and Walker, S., "Fitting linear mixed-effects models using lme4", *Journal of Statistical Software,* vol. 67, no. 1, 1-48, 2015, doi: 10.18637/jss.v067.i01.

[31] *ggplot2: Elegant graphics for data analysis*. (2016). Springer-Verlag, New York, NY. [Online]. Available: http://ggplot2.org

[32] *emmeans: Estimated Marginal Means, aka Least-Squares Means*. (2023). [Online]. Available: https://CRAN.R-project.org/package=emmeans

[33] Jepson, K., "Prosody, prominence and segments in Djambarrpuyŋu", PhD thesis, University of Melbourne, Melbourne, Australia, 2019.

An Acoustic and Electroglottographic (EGG) Investigation of Preaspiration and Voice Quality in the Italian Four-Way Stop Contrast across Regional Accents

Angelo Dian, John Hajek, Janet Fletcher

School of Languages and Linguistics, The University of Melbourne

a.dian@unimelb.edu.au; j.hajek@unimelb.edu.au; j.fletcher@unimelb.edu.au

Abstract

This preliminary study explores the effects of consonant voicing and gemination on preaspiration and voice quality patterns of stops in Italian, a language that features considerable cross-regional variability in the phonetic realization of its four-way stop contrast. Five speakers, each from a different Italian region, produced /maC(:)a/ words. Findings of an acoustic and dynamic electroglottographic (EGG) analysis reveal that voiceless stops consistently lead to a breathier voice quality in the final 30% of the preceding vowel, while there is no overall effect of gemination on vowel voice quality. Voiceless preaspiration varies by region: Northern speakers preaspirate both voiceless geminates and singletons frequently, while Centro-Southern speakers preaspirate geminates more often. The study concludes that breathy voice preceding voiceless stops may be characteristic in Italian, while preaspiration patterns are influenced by individual, articulatory, and possibly regional factors.

Index Terms: preaspiration, voice quality, EGG, Italian stop contrast, regional variation.

1. Introduction

1.1. Background

Italian is a well-known example of a language exhibiting a four-way stop consonant contrast along two dimensions, voicing and length, as exemplified by the minimal set /rita rida rit:a rid:a/ ‘Rita (given name), laugh (pres. subj. 123s), upright (f), ridda (type of dance)’. Voicing and length distinctions find their phonetic expression primarily through the vibration of the vocal folds and the duration of the consonantal gesture, respectively, suggesting a complex interplay between glottal and supraglottal articulations in the implementation of the four-way contrast. Furthermore, substantial variation exists in the realization of this contrast across regional “accents”, as outlined below.

1.1.1. Stop voicing and voice quality

Italian features a series of voiced /b b: d d: g g:/ and voiceless stops /p p: t t: k k:/, contrasting singleton (short) and geminate (long) phonemes. Voiced stops are typically produced with active vocal fold vibration, or ‘prevoicing’ [1], during closure in all contexts across varieties (although the degree of prevoicing especially in geminates may vary cross-regionally [2]). On the other hand, voiceless stop realizations may differ by variety. While voiceless geminate /p: t: k:/ are normally realized as voiceless [p: t: k:] (that is, with little or no vocal fold vibration during closure) cross-regionally, intervocalic singleton /p t k/ can surface either: (a) as also phonetically voiceless [p t k] in the North; (b) with optional prevoicing, e.g. [p̌ t̬ k̬]/[β ð ɣ], in the Centre-South; or (c) as typically voiceless fricatives [ɸ θ h] (but also /k/ > [ɦ] [3]) in Tuscan varieties, a phenomenon known as *gorgia toscana*, involving intervocalic singleton stop spirantization [3], [4]). To actively implement phonetic voicelessness of consonants in a post-vocalic context, speakers can either spread or constrict the vocal folds [5]. The glottal spreading gesture can result in a breathier voice quality in the preceding vowel, which may be accompanied by preaspiration, i.e., a period of glottal friction preceding the onset of supraglottal constriction [6]. The glottal constriction gesture may instead result in a creakier voice quality and concomitant preglottalization. In fact, some preglottalized voiceless stops, alongside more frequent preaspirated ones, have been observed acoustically as separate allophones in Italian [7].

1.1.2. Gemination and voice quality

Italian geminates, as compared to singletons, are cued durationally by: (i) a longer consonant (C) duration (~2x longer in Central varieties [8]-[10]); (ii) a shorter duration of preceding stressed vowels (V) (between ~20-50% shorter across varieties [2], [11]); and (iii) a higher ratio of consonant to preceding-vowel duration (C/V) [2], [12]. In Northern varieties, however, the difference between geminate and singleton C duration and C/V ratio may be reduced due to a phonetic lengthening of voiceless singletons [2], [13] and potential (although not yet proven) geminate shortening, e.g. [14].

Importantly, it has also been shown that singletons and geminates differ not only in acoustic durational properties, but also spatiotemporally. Specifically, Italian geminates, in comparison to singletons, are more constricted [15], [16], are produced with a higher tongue position when lingual [17], and exhibit an earlier initiation of the consonantal gesture relative to the tongue gesture in preceding V [16]. These varying tongue adjustments may result in laryngeal modifications through indirect movements of the tongue root (which is connected to the larynx via the hyoid bone [18]) ultimately affecting voice quality (cf. [19], [20]). On this point, preaspiration, in turn linked to a breathier voice quality in preceding V, has been found in voiceless geminates across Italian regional varieties [21]. Recently, however, it has been demonstrated that preaspiration of Italian singletons may also occur, although its frequency of occurrence may vary cross-regionally, with Centro-Southern varieties showing it less frequently than Northern varieties ([22]; also cf. [23] for a Tuscan variety).

As yet, no previous work has specifically looked at V voice quality patterns in connection with the Italian C length contrast. Moreover, previous studies on voice quality associated with gemination in other languages have mostly focussed on the following vowel. For example, geminate stops in Japanese [24] and Lebanese Arabic [25] have been associated with a creakier voice quality following the release, while in, e.g., Cypriot Greek they have been linked to breathy voice [26], suggesting that voice quality patterns around geminates may be language

Table of Contents for this manuscript

specific. Furthermore, these previous studies have tended to look at acoustic measures of voice quality exclusively.

1.2. Aims

This preliminary study is the first to examine voice quality through electroglottography (EGG) in relation to both the C voicing and length contrasts in Italian, with two primary aims. First, it seeks to establish a connection between speaker-specific tendencies of preaspiration and/or preglottalization occurrence and breathiness/creakiness in V preceding voiceless stops across consonant length categories. The second aim has a broader scope, namely, to explore the effects of consonant voicing and gemination on V voice quality in Italian.

2. Methods

2.1. Participants

Five adult speakers took part in the study (see Table 1), each from a different city of Italy across three broader regions, according to [27]'s classification. Two participants came from the North (N): N_Tur_F, N_Vic_M; one from Tuscany (T): T_Emp_M; and two from the Centre-South (CS): CS_Rom_F, CS_Cat_M. All participants were born, raised, and resided most of their lives in their city/region of origin, although they had lived in Melbourne, Australia, for a period not exceeding 2.5 years at the time of the study. They reported daily usage of Italian. Details regarding the age, sex, city of origin, and length of stay in Melbourne for each speaker are provided in Table 1.

Table 1: *Participant details.*

Speaker ID	Age	Sex	City	Length of stay (years)
N_Tur_F	37	F	Turin	0.25
N_Vic_M	37	M	Vicenza	2.5
T_Emp_M	19	M	Empoli	0.16
CS_Rom_F	43	F	Rome	2.5
CS_Cat_M	24	M	Catania	2.5

2.2. Materials and procedure

An acoustic and articulatory experiment using EGG was designed for the study. EGG is a non-invasive technique that measures the contact area between the vocal folds, providing information on vocal fold activity at the source, prior to modification by the supralaryngeal articulators [28].

Participants were instructed to produce a series of (mostly) nonce /ˈmaC(ː)a/ words, where C(ː) represents all Italian oral stop phonemes (cf. §1.1.1) as well as nasal /m mː/, used as the baseline for the EGG data (see §3.2). Note that all resulting words adhere to Italian phonotactic rules and are plausible real words in the language. In fact, at least four of the 14 total words, namely /ˈmapːa ˈmatːa ˈmaga ˈmamːa/, meaning 'map', 'mad (f)', 'magician (f)', and 'mum', are commonly employed in Italian. This particular phonetic structure was selected to avoid the confounding effect of different word-initial C on glottal cycle characteristics in the following V, as noted in initial trials using real words. Instead, employing word-initial /m/ for all tokens triggered the smallest amount of glottal cycle perturbation in the following V. Additionally, /a/ was chosen because preaspiration occurs most commonly in low vowels [21], [29].

The 14 experimental words were embedded in the Italian carrier phrase "*Dico WORD lentamente*" 'I say WORD slowly' and read out five times by all 5 participants, resulting in 350 total tokens. The phrases were displayed through a PowerPoint presentation one by one in random order.

2.3. Data collection and analysis

An EGG-D400 Laryngograph was used to collect the data. The acoustic signal was captured and synchronized with the EGG signal through a RØDE NT3 microphone connected to the Laryngograph, set at a 48 kHz sampling rate. The EGG signal was obtained through a pair of electrodes placed on each side of the participants' thyroid cartilage. The recordings were conducted in a quiet room at the University of Melbourne.

The combined acoustic and EGG signals generated by the Laryngograph were analyzed in Praat. The acoustic signal was force-aligned using WebMAUS [30] to obtain phonetic annotations, and target V boundaries were subsequently adjusted manually where necessary. Boundaries were placed based on the acoustic signal alone, at the onset and offset of vowel-like periodicity and formant structures (see Figure 1). Additionally, voiceless preaspiration of voiceless stops, visible as diffuse aperiodic energy observed in the 0-12 kHz range preceding stop closure, was also annotated and its frequency of occurrence counted.

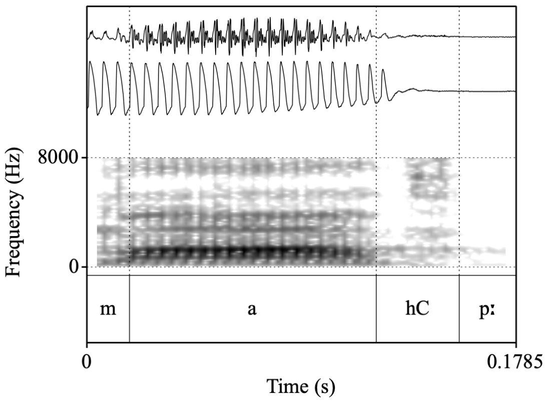


Figure 1: *Annotated example of a /mapːa/ token produced by N_Tur_F. The acoustic waveform is at the top and the synchronized EGG signal is underneath it. The acoustic spectrogram is at the bottom, revealing the presence of voiceless preaspiration, labelled 'hC'.*

The EGG signal was processed using the Praatdet script [31], which extracted open quotient (OQ) measurements across V duration. Following [32], Howard's OQ measure was chosen for the present investigation. OQ is a measure of glottal spreading/constriction and is defined as the duration of glottal opening over the duration of the entire glottal cycle [33]. Hence, the higher the OQ, the breathier (or less creaky) the voice.

2.4. Statistical analysis

For comparability across speakers, the time-aligned OQ data underwent z-score normalization by speaker. Two Generalized Additive Mixed Models (GAMMs) were fitted through the mgcv [34] and itsadug [35] packages in R [36]. GAMM 1 included (a) a parametric (P) interaction between speaker (5 levels) and C type (6 levels: singleton nasal, geminate nasal, voiced singleton stop, voiced geminate stop, voiceless singleton stop, and voiceless geminate stop), resulting in 30 interaction levels, and (b) a smooth (S) term over normalized V duration by the interaction of speaker and C type. Singleton nasals produced by N_Tur_F were set as the baseline (cf. §3.2). GAMM 2 focussed on stops and had voicing (voiced/voiceless) and gemination (geminate/singleton) as parametric (P) terms, with two smooth (S) terms over normalized V duration, the first by voicing and the second by gemination. Both GAMMs included a random smooth term over normalized V duration by speaker. Basis functions were set to ten (k = 11). An AR1 error term was included to address autocorrelation. Post-hoc tests were conducted on parametric differences using the emmeans function and package [37].

3. Results

3.1. Preaspiration/preglottalization patterns

Preaspiration of voiceless stops occurred frequently (91/150, 61% of total tokens), while, unlike [7], no instances of preglottalization were found. Table 2 provides counts of preaspirated tokens by participant and phoneme, showing that individual participants varied considerably in rates of preaspiration occurrence. N_Tur_F preaspirated nearly all voiceless stops (28/30 tokens) regardless of length. N_Vic_M preaspirated /t t: k k:/ most frequently (18/20). T_Emp_M only preaspirated geminates (13/15 – but see below). CS_Rom_F also showed a preference for geminate preaspiration (14/15), although singletons were also sometimes preaspirated (5/15). Finally, CS_Cat_M mostly preaspirated velars (5/5 /k:/ and 4/5 /k/), but not other places of articulation (except for 2/5 /t:/ tokens). A qualitative inspection of the corpus showed that CS_Rom_F and CS_Cat_M also tended to exhibit some prevoicing at the onset of /p t/ closure (see more in §3.2).

Region-specific patterns can also be observed. N speakers produced preaspiration more frequently (48/60, or 80% of total tokens) than the T and CS speakers (13/30, or 43% of total tokens for T and 30/60, or 50% of total tokens for CS speakers). This is because N speakers frequently preaspirated singletons as well as geminates (22/30, or 73% preaspirated singletons for N speakers as compared to no occurrences for the T speaker and 9/30, or 30% occurrences for CS speakers).

Table 2: *Counts of tokens exhibiting voiceless preaspiration.*

	/p/	/t/	/k/	/p:/	/t:/	/k:/	Tot
N_Tur_F	5/5	4/5	5/5	5/5	5/5	4/5	28/30
N_Vic_M	0/5	4/5	4/5	2/5	5/5	5/5	20/30
T_Emp_M	0/5*	0/5*	0/5*	4/5	5/5	4/5	13/30
CS_Rom_F	0/5	2/5	3/5	4/5	5/5	5/5	19/30
CS_Cat_M	0/5	0/5	4/5	0/5	2/5	5/5	11/30
All	5/25	10/25	16/25	15/25	22/25	23/25	91/150

Note. *These tokens were all spirantized as [ɸ θ ɦ]

It should be noted that T_Emp_M consistently spirantized /p t k/ as [ɸ θ ɦ], as previously reported for Tuscan Italian [3] (see examples in Figure 2). Preaspiration occurring as voiceless glottal friction, similar to Figure 1, was not found in any of these spirantized tokens (in contrast to [23]).

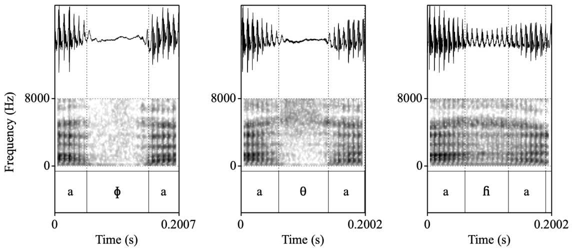


Figure 2: *Annotated examples of /mapa mata maka/ showing intervocalic spirantization produced by T_Emp_M.*

Following the categorical, qualitative analysis of preaspiration occurrence above, we now proceed to a quantitative analysis of V voice quality patterns to examine potential gradience in the voice quality continuum.

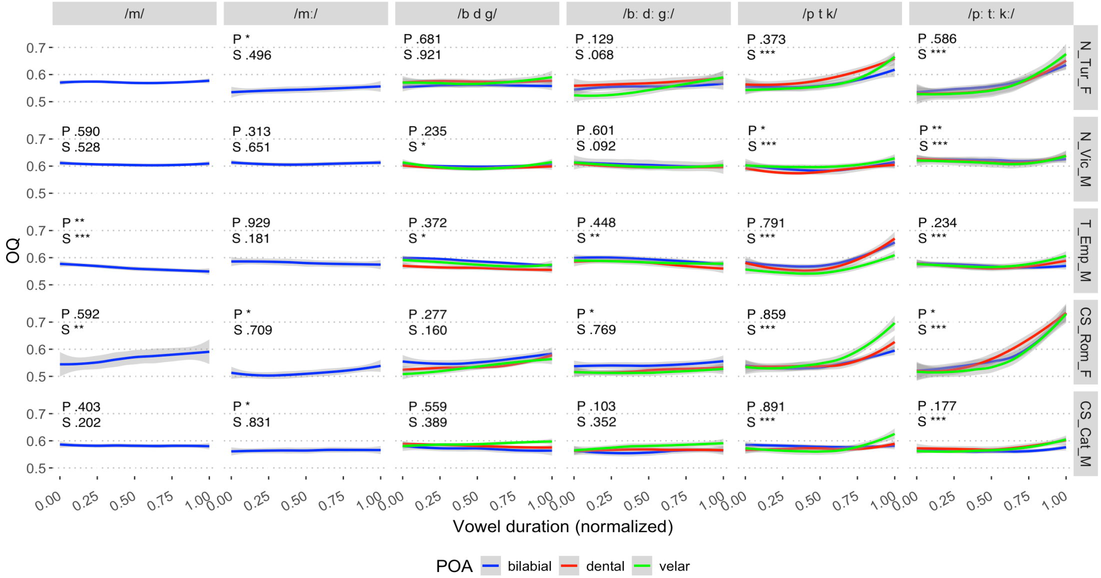


Figure 3: *Mean trajectories of open quotient (OQ) across normalized durations of preceding vowels by consonant type and place of articulation (POA) for each speaker.*

Table of Contents for this manuscript

3.2. Voice quality in the preceding vowel

Figure 3 displays average OQ trajectories across the duration of V preceding target C types by place of articulation (POA) for each speaker. The same figure shows results of GAMM 1, reporting significance levels of parametric (P) and non-linear smooth (S) differences in the z-score data compared to the baseline /m/ trajectory produced by N_Tur_F. Significant differences are marked with asterisks, indicating p < .05 or lower, while actual p-values are provided for non-significant differences. Note that the baseline trajectory serves as an ideal reference for other trajectories, as it is nearly a straight horizontal line centered at 0 in the z-score data.

Overall, OQ trajectories differ substantially in shape across voiceless and voiced stop tokens, as signalled by the non-linear smooth term ‘S’. For V preceding singleton and geminate voiceless stops, they have a non-flat shape, all at p < .001 – specifically, they are rising for all speakers, indicating a progressively breathier voice towards V offset. By contrast, OQ in V preceding voiced stops tends to stay level throughout V duration across speakers, with two exceptions: (i) for T_Emp_M, OQ decreases slightly towards V offset preceding both singletons and geminates, pointing to creakier voice; (ii) N_Vic_M, exhibits a somewhat falling-rising OQ contour preceding singletons. Another thing to note is the speaker-specific variation in voice quality preceding /m/, with a OQ trajectory that is falling for T_Emp_M and rising for CS_Rom_F, highlighting speaker-specific voice quality patterns even in the supposedly more neutral /mVm/ context.

Parametric differences from the baseline, denoted by ‘P’, were also detected. For voiceless stops, overall higher OQ values were observed in N_Vic_M for both singletons and geminates, and in CS_Rom_F exclusively for geminates. For voiced stops, lower OQ values preceding geminates are displayed by CS_Rom_F alone. For nasals, lower OQ preceding geminate /m:/ was detected in three speakers: N_Tur_F, CS_Rom_F, and CS_Cat_M. By contrast, only T_Emp_M exhibited lower OQ than other speakers preceding singleton /m/.

Post-hoc tests were run to investigate the effect of gemination on OQ trajectories in preceding V for each individual speaker across voiced and voiceless stop series. There was only one significant parametric difference between singleton and geminate voiceless stops for N_Vic_M (β = -1.325, SE = 0.211, t = -6.295, p < .001), with V preceding /p: t: k:/ exhibiting higher OQ than V preceding /p t k/ for this speaker. Voiced stops and nasals did not show significant singleton-geminate differences for any of the speakers.

GAMM 2 tested the overall effects of stop voicing and gemination on the OQ trajectories across speakers. Parametric differences were statistically significant between voiceless and voiced stops, at p < .01, but not between geminate and singleton stops, at p = 0.784. Non-linear smooth differences were significant for both voicing and gemination, both at p < .001. Figure 4 illustrates that OQ is slightly lower for V preceding voiceless stops up to ~50% of V duration, with this trend reversing from ~70% of V duration where OQ becomes higher for voiceless stops with this difference increasing steeply towards V offset.

Some variation in OQ trajectories specific to the place of articulation of the following stop can be observed in Figure 3, although this was not tested statistically for lack of space here. Generally, it appears that V preceding velars has a higher OQ than preceding non-velars towards its offset in the case of singletons produced by the CS speakers, while the T speaker shows the opposite trend (note, however, that this speaker produced /k/ as a voiced [ɦ] in all cases). This phenomenon may be due to an observed tendency for the CS speakers to produce /p t/, but not /k/, with some vocal fold vibration in the first ~20% of stop closure duration. This observation is mirrored by the low rate of occurrence of preaspiration of /p t/ for these speakers reported in Table 2.

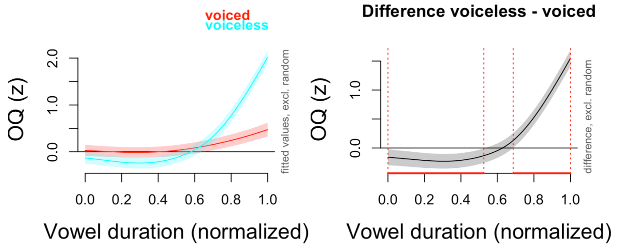


Figure 4: *Predicted smooths (left) and smooth differences (right) of OQ (z) for voiced and voiceless stops over normalized preceding-vowel duration.*

4. Discussion and conclusion

This study finds that voiceless stops, as compared to voiced stops, and regardless of phonological length, are characterized by a breathier voice quality (higher OQ) in approximately the final 30% of preceding-V duration across regional varieties of Italian. This finding aligns with and adds to [38] who looked exclusively at singletons. Moreover, increased creakiness (lower OQ) or preglottalization never occurred before voiceless tokens in the present data, although somewhat creakier voice surfaced preceding voiced tokens in some cases.

Another key finding is that C gemination does not have an overall effect on the voice quality of preceding V, although one Northern speaker (N_Vic_M) did produce increased breathiness preceding voiceless geminates vs. singletons. Furthermore, the occurrence of voiceless preaspiration appears to be partly speaker- and potentially region-specific. In our data, Northern speakers exhibit comparable rates of preaspiration for voiceless geminates and singletons, whereas Centro-Southern speakers show a higher propensity to preaspirate voiceless geminates, corroborating findings from [22]. However, data from more speakers for each region are needed in future studies to support this generalization. Additionally, the lower OQ observed preceding tendentially pre-voiced singleton /p t/ produced by Centro-Southern speakers suggests that these speakers may not spread the vocal folds before voiceless singletons as actively, or to the same extent, as they do for voiceless geminates. Conversely, Northern speakers actively devoice all voiceless stops (cf. [2], [22]). Interestingly, this devoicing is always implemented through glottal spreading in this study, and never through glottalization as observed in some cases by [7]. As for the Tuscan speaker, who regularly spirantized /p t k/ intervocalically, breathy voice in preceding V was more pronounced for singletons than geminates, supporting the notion that (voiceless) fricatives typically exhibit substantial breathiness in the V-C transition [39].

In conclusion, this study suggests that breathy voice towards V offset, indicative of glottal abduction, is a gradient phenomenon that may be characteristic of voiceless stops in Italian while voiceless preaspiration occurs more variably, influenced by factors such as: (a) speaker-specific tendencies, (b) C length, (c) C place of articulation, and possibly (d) regional pronunciation.

5. References

[1] A. S. Abramson and D. H. Whalen, ‘Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions’, *Journal of Phonetics*, vol. 63, pp. 75–86, 2017, doi: 10.1016/j.wocn.2017.05.002.

[2] A. Dian, J. Hajek, and J. Fletcher, ‘Cross-regional patterns of obstruent voicing and gemination: The case of Roman and Veneto Italian’, *Languages*, under review.

[3] G. Marotta, ‘Lenition in Tuscan Italian (Gorgia Toscana)’, in *Lenition and fortition*, J. Brandão de Carvalho, T. Scheer, and P. Ségéral, Eds., Berlin: Mouton de Gruyter, 2008, pp. 235–272.

[4] P. M. Bertinetto and M. Loporcaro, ‘The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome’, *Journal of the International Phonetic Association*, vol. 35, no. 2, pp. 131–151, 2005, doi: 10.1017/S0025100305002148.

[5] M. Garellek, ‘The phonetics of voice’, in *The Routledge Handbook of Phonetics*, 1st ed., W. F. Katz and P. F. Assmann, Eds., Abingdon, Oxon; New York, NY: Routledge, 2019, pp. 75–106. doi: 10.4324/9780429056253-5.

[6] P. Helgason, ‘Preaspiration in the Nordic Languages: Synchronic and diachronic aspects’, PhD dissertation, University of Stockholm, 2002.

[7] M. Stevens and J. Hajek, ‘Towards a phonetic conspectus of preaspiration’, in *Proc. 16th ICPhS*, Saarbrücken, 2007, pp. 429–432.

[8] P. Mairano and V. De Iacovo, ‘Gemination in Northern versus Central and Southern varieties of Italian: A corpus-based investigation’, *Lang Speech*, vol. 63, no. 3, pp. 608–634, 2020, doi: 10.1177/0023830919875481.

[9] E. M. Payne, ‘Phonetic variation in Italian consonant gemination’, *Journal of the International Phonetic Association*, vol. 35, no. 2, pp. 153–181, 2005, doi: 10.1017/S0025100305002240.

[10] A. Esposito and M. G. Di Benedetto, ‘Acoustical and perceptual study of gemination in Italian stops’, *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 2051–2062, 1999, doi: 10.1121/1.428056.

[11] J. Hajek and M. Stevens, ‘Vowel duration, compression and lengthening in stressed syllables in Central and Southern varieties of Standard Italian’, *Proc. Interspeech 2008*, Brisbane, pp. 516–519, 2008.

[12] E. R. Pickett, S. E. Blumstein, and M. W. Burton, ‘Effects of speaking rate on the singleton/geminate consonant contrast in Italian’, *Phonetica*, vol. 56, no. 3–4, pp. 135–157, 1999, doi: 10.1159/000028448.

[13] L. Canepari, *Lingua italiana nel Veneto*. Padua: CLESP, 1984.

[14] L. Canepari, *Manuale di pronuncia italiana*, 1. ed. Bologna: Zanichelli, 1992.

[15] E. M. Payne, ‘Non-durational indices in Italian geminate consonants’, *Journal of the International Phonetic Association*, vol. 36, no. 1, pp. 83–95, 2006, doi: 10.1017/S0025100306002398.

[16] F. Burroni, S. Maspong, N. Benker, P. Hoole, and J. Kirby, ‘Spatiotemporal features of bilabial geminate and singleton consonants in Italian’, in *Proc. 13th ISSP,* Autrans, 2024.

[17] D. Dipino and C. Celata, ‘An UTI study of alveolar stops in Italian’, *Il parlato nel contesto naturale. Speech in the natural context*, no. 4, pp. 41–53, 2018, doi: 10.17469/O2104AISV000003.

[18] R. C. Auvenshine and N. J. Pettit, ‘The hyoid bone: an overview’, *CRANIO®*, vol. 38, no. 1, pp. 6–14, 2020, doi: 10.1080/08869634.2018.1487501.

[19] J. Kingston, N. A. Macmillan, L. W. Dickey, R. Thorburn, and C. Bartels, ‘Integrality in the perception of tongue root position and voice quality in vowels’, *The Journal of the Acoustical Society of America*, vol. 101, no. 3, pp. 1696–1709, 1997, doi: 10.1121/1.418179.

[20] S. G. Guion, M. W. Post, and D. L. Payne, ‘Phonetic correlates of tongue root vowel contrasts in Maa’, *Journal of Phonetics*, vol. 32, no. 4, pp. 517–542, 2004, doi: 10.1016/j.wocn.2004.04.002.

[21] M. Stevens, ‘How widespread is preaspiration in Italy?’, *Working Papers of the Department of Linguistics and Phonetics 2010, Lund University*, vol. 54, pp. 97–102, 2010.

[22] A. Dian, J. Hajek, and J. Fletcher, ‘Preaspiration in Italian voiceless geminate and singleton stops’, in *Proc. 20th ICPhS*, Prague, 2023, pp. 888–892. doi: 10.13140/RG.2.2.25014.88648.

[23] M. Stevens and J. Hajek, ‘Spirantization of /p t k/ in Sienese Italian and so-called semi-fricatives’, in *Proc. Interspeech 2005*, Lisbon, 2005, pp. 2893–2896.

[24] K. Idemaru and S. G. Guion, ‘Acoustic covariants of length contrast in Japanese stops’, *Journal of the International Phonetic Association*, vol. 38, no. 2, pp. 167–186, 2008, doi: 10.1017/S0025100308003459.

[25] J. Al-Tamimi and G. Khattab, ‘Acoustic correlates of the voicing contrast in Lebanese Arabic singleton and geminate stops’, *Journal of Phonetics*, vol. 71, pp. 306–325, 2018, doi: 10.1016/j.wocn.2018.09.010.

[26] A. Arvaniti and G. Tserdanelis, ‘On the phonetics of geminates: evidence from Cypriot Greek’, in *Proc. 6th ICSLP*, Beijing, 2000, pp. 559-562.

[27] G. B. Pellegrini, *Carta dei dialetti d'Italia*. Pisa: Pacini.

[28] M. Fabre, ‘Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation: glottographie de haute fréquence. Premiers resultats.’, *Bulletin de l'Académie Nationale de Médecine*, vol. 141, no. 66, 1957.

[29] M. Hejná, ‘Pre-aspiration in Welsh English: A case study of Aberystwyth’, PhD dissertation, University of Manchester, 2015. Available: http://rgdoi.net/10.13140/RG.2.1.3485.3842

[30] F. Schiel, ‘A statistical model for predicting pronunciation’, in *Proc. 18th ICPhS,* Glasgow, 2015.

[31] J. Kirby, ‘Praatdet: Praat-based tools for EGG analysis’. 2020. [Online]. Available: https://github.com/kirbyj/praatdet

[32] L. Ratko, J. Penney, and F. Cox, ‘Opening or closing? An electroglottographic analysis of voiceless coda consonants in Australian English’, in *Proc. Interspeech 2023*, Dublin, 2023, pp. 1823–1827. doi: 10.21437/Interspeech.2023-2337.

[33] H. M. Hanson, K. N. Stevens, H.-K. J. Kuo, M. Y. Chen, and J. Slifka, ‘Towards models of phonation’, *Journal of Phonetics*, vol. 29, no. 4, pp. 451–480, 2001, doi: 10.1006/jpho.2001.0146.

[34] S. N. Wood, ‘Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models’, *Journal of the Royal Statistical Society (B)*, vol. 73, no. 1, pp. 3–36, 2011.

[35] J. van Rij, M. Wieling, R. H. Bayeen, and H. van Rijn, ‘itsadug: Interpreting time series and autocorrelated data using GAMMs’. 2022.

[36] R Core Team, ‘R: A language and environment for statistical computing’. R foundation for statistical computing, Vienna, Austria, 2023.

[37] R. V. Lenth *et al.*, ‘Package “emmeans”’. 2022. [Online]. Available: https://cran.r-project.org/web/packages/emmeans/emmeans.pdf

[38] S. Coretta, ‘Vowel duration and consonant voicing: A production study’, PhD dissertation, University of Manchester, 2020. Available: https://www.research.manchester.ac.uk/portal/files/173354658/FULL_TEXT.PDF

[39] C. Gobl and A. Ní Chasaide, ‘Voice source variation in the vowel as a function of consonantal context’, in *Coarticulation: Theory, data and techniques*, W. J. Hardcastle and N. Hewlett, Eds. Cambridge: Cambridge University Press, 2000, pp. 122–143.

Fillers and Creaky Voice Presence in Australian English

Hannah White, Joshua Penney, Felicity Cox

Department of Linguistics, Macquarie University, Australia
hannah.white@mq.edu.au; joshua.penney@mq.edu.au; felicity.cox@mq.edu.au

Abstract

The fillers *um* and *uh* have been shown to vary in terms of both their linguistic functions and their acoustics, including voice quality. The present study explores how, in Australian English, the likelihood of creaky voice presence differs depending on the phonetic realisation of the filler (*um* vs *uh*) and whether a token is a filler or a lexical item. Results suggest that different fillers do vary in their acoustics, and that social factors do not affect creaky voice presence on fillers in the same way that they do lexical items.

Index Terms: fillers, creaky voice, Australian English

1. Introduction

Fillers such as *um* and *uh* are non-lexical vocalisations, which are used widely cross-linguistically [1]. Fillers have been proposed to hold several functions in English, including planning upcoming speech, signalling upcoming delays and holding/ceding the floor [1, 2, 3], as well as the social meaning of uncertainty [3, 4].

Rates of filler use as well as preference for either *um* or *uh* vary according to macro-social categories such as gender, age and socioeconomic status [5, 6, 7, 8, 9, 10]. Across British English (BrE) [6, 7] and American English (AmE) [9] more fillers have been reported in men's speech compared to women's speech (although see [5]). Comparing *um* to *uh* across different Germanic languages, women and young speakers have been shown to prefer *um* to *uh* more than men and older speakers [6, 7, 8, 10], perhaps indicating a cross-linguistic change in progress towards *um* use as suggested by [8] and [10].

From a phonetic perspective, fillers have been investigated as potentially useful for forensic speaker comparisons as they are generally produced unconsciously and therefore less likely to be affected by a speaker's attempt to disguise their voice [11]. [11] examined formant frequencies and duration of *um* and *uh* for usefulness in forensic speaker discrimination in male speakers of BrE. Using the likelihood ratio framework, they determined that fillers were promising candidates for forensic speaker comparison when the first three formants and duration measurements were incorporated [11]. In a study of bilingual English-Māori speakers in New Zealand, [12] found that *um* and *uh* were speaker-specific in their formant and fundamental frequency (f0) characteristics with *um* discriminating between speakers more accurately compared to *uh*.

Language context can affect filler production. [13] examined the acoustics of *um* and *uh* in female L1 Dutch speakers learning English as L2. Both F1 and F2 differed between L1 Dutch and L2 English for both fillers; however, neither *um* nor *uh* differed in duration or f0 across the two language contexts [13]. Position in utterance influenced f0 of fillers with both *um* and *uh* having significantly higher f0 at the beginnings of utterances compared to in utterance-medial or final position or when surrounded by silence [13]. [14] examined the F1, F2, duration and f0 of fillers from eight different languages including Arabic, Mandarin Chinese, French and AmE. They found that while duration and f0 were relatively stable across languages, F1 and F2 frequencies were language-specific.

Many more studies have investigated the pitch characteristics of fillers across languages such as Spanish, German, French, Japanese and English [15, 16, 14, 17, 18, 19, 20]. For English, [18] found that fillers tended to have falling or level f0 contours and that their end points occurred within the lowest 15% of a speaker's f0 range regardless of whether they occurred at a phrase boundary or medially. [19, 20] provided evidence that the f0 of clause-internal fillers can be predicted by the f0 of the surrounding prosodic context, with fillers almost always having lower f0 compared to their surroundings. Position can also impact f0 with [21] finding that for both *um* and *uh* in Dutch, tokens in utterance-initial position had significantly higher f0 and longer duration than medial tokens (similar to [13]'s cross-linguistic findings). Studies have also shown differences between *um* and *uh* in Dutch, with *um* having higher mean f0 and longer mean durations [21].

More recently, studies have examined voice quality characteristics of fillers [22, 23]. [22] compared measures of voice quality between vowels in Japanese fillers and corresponding vowels in lexical items. For males, fillers had lower mean f0 compared to lexical items; however, there was no difference for females. For both males and females, fillers had higher H1-H2 ratios (correlated with a breathier voice quality) and higher jitter and shimmer (indicating more irregularity/aperiodicity in the signal) [22]. Similar studies comparing Chinese and Japanese showed cross-linguistic differences with lower jitter and shimmer in fillers compared to lexical items for Chinese speakers [24, 23]. Both [22] and [23] found that voice quality measures were important cues to classifying fillers versus lexical vowels. Studies of French and English have noted that fillers tend to be produced with creaky voice quality [9, 14]. Creaky voice (or creak) is a voice quality generally characterised by low f0 and irregular/aperiodic glottal pulses [25]. In French, [14] found that Praat with default settings failed to detect f0 more often in fillers compared to lexical items, particularly for men, which they attribute to "an unstable voice quality which could be either vocal fry, creaky or breathy". In English, fillers have been anecdotally noted to frequently co-occur with creak [9]. [9] points out that disfluencies such as fillers can cause problems for automatic speech recognition (ASR) technology. Increased knowledge of the acoustic correlates of fillers could lead to improvements of ASR algorithms [9].

To our knowledge, the phonetic characteristics of fillers in Australian English (AusE) is yet to be examined. The aim of the present study is to explore the relationship between creak and fillers in AusE through two analyses. The first explores whether there are differences between different fillers by comparing creak occurrence between *um* (/ɐm/) and *uh* (/ɐː/). The second compares creak presence between *um* and lexical items in the same phonetic context (i.e., /ɐ/+nasal). Creak can signal phrase/turn-finality across a number of languages, including AusE [26, 27, 28, 29]. In AusE, [30] found that male speakers of Aboriginal AusE produced creakier voice quality than non-indigenous males, while [31, 32] have shown that creak use in AusE adolescents varies according to speaker gender and language background, with speakers from a monolingual English background using more creak than those with at least one parent with a first language other than English (LOTE). Among English-only background speakers, female speakers were found to use more creak than male speakers with no gender difference among LOTE background speakers [31, 32]. As previous studies have shown the acoustic characteristics of fillers may differ according to gender and language context, as well as position in utterance, these variables are included in the present analyses [14, 13, 22, 23].

2. Methods

2.1. Data

The data used in this study consist of conversations between 131 AusE-speaking teenagers from different areas of Sydney extracted from the Multicultural Australian English - Voices of Sydney (MAE-VoiS) corpus [33]. One speaker identified as neither male or female and, due to a lack of statistical power, was excluded from analysis resulting in the inclusion of 130 speakers (61f; 69m). Speakers were aged between 15 and 19 years (mean age = 15.7). 27 speakers were from monolingual English households (16f; 11m) with the remaining from LOTE backgrounds (45f; 58m). Recordings ranged in length from 5 to 35 minutes (mean length = 16 minutes; total length = 20 hours). Conversations were guided by a trained research assistant (RA). All but four of the included speakers conversed with a partner matched for gender. Speakers were generally familiar with their conversational partner (i.e., they were in the same class at school) apart from in 13 cases where the speaker conversed only with the RA.

Creaky voice was identified using the automatic optimised Union method [34]. This method combines an approach based on identifying creak through low f0 (the AntiMode method) [35, 36] with one that uses other acoustic cues to creak such as a measure of spectral tilt and residual peak prominence (the Creak Detector algorithm) [37] and has been shown to significantly improve creak identification compared to when each tool is used on its own [34]. The Union method identifies anything coded by either AntiMode or Creak Detector as creaky voice and returns a binary creak decision for every 10 ms of speech data.

All conversations were automatically transcribed using the IBM Watson Speech to Text API (https://www.ibm.com/cloud/watson-speech-to-text). Orthographic transcripts were then manually corrected by RAs and checked by trained phoneticians. Final transcripts were processed through the MAUS automatic forced-aligner [38].

While automatic detection methods are useful for annotating large quantities of data, they are not infallible. For this reason, phoneme boundaries of *um*, *uh* and all lexical /ɐ/+nasal tokens were manually corrected. The Union method was then checked within these tokens, with tokens coded as either creaky (if they contained creak) or not creaky. In tokens where the vowel was word-initial, creak intervals of 50 ms or less at the onset of the segment were considered to be segmental glottalisation (for tokens with an overall duration less than 100 ms, segmental glottalisation was considered present if less than half the vowel contained creak) [40, 41]. Segmental glottalisation in initial position has been proposed as a strategy of prosodic strengthening in English and differs from phrasal creak, which can be linked to conveying social meaning [26, 31, 41]. For this reason, and the fact that it is often too brief to be identified by the Union method, tokens that only contained segmental glottalisation and no other creak were coded as not creaky.

For each token of *um*, *uh* and lexical /ɐ/+nasal, position in the utterance was manually coded. This process was guided by the method used in [13]. Tokens were coded as 'isolated' if they were the speaker's entire conversational turn. If they occurred at the end of a phonological phrase, they were coded as 'phrase-final'. Similarly, if they occurred at the beginning of a phonological phrase, they were coded as 'phrase-initial'. Finally, tokens were coded as 'phrase-medial' if they interrupted a phonological phrase. Following [14], tokens were excluded if they were shorter than 40 ms (0.3% of all tokens). The duration of each token (i.e., vowel for *uh*, vowel + nasal for *um* and lexical items) were extracted in milliseconds.

2.2. Analysis

Two analyses were carried out. These were conducted with generalised linear mixed effects regression (GLMER) modelling using the lme4 [42] and lmerTest [43] packages in R [44]. The first analysis compared creak presence in *um* versus *uh*. The dependent variable was whether creak was present or not. A model was built with fixed effects of word (*um* or *uh*), speaker gender, whether the speaker had a LOTE background or not (henceforth referred to as LOTE-BG), position in utterance and duration. Three-way interactions were included for word, gender and LOTE-BG and word, position and duration. Duration was *z*-scored. A random intercept was included for speaker with random slopes for word, position and duration. The full model returned a singular fit warning so a stepwise reduction approach was taken, systematically reducing non-significant interactions followed by non-significant fixed effects and random slopes, each time comparing models using ANOVAs to ensure the most parsimonious model. The final model syntax is shown in 1. 2491 tokens (*um*=1787; *uh*=704) were included.

$$\text{Creak} \sim \text{word} + \text{pos} + \text{dur} + (1 + \text{word} \mid \text{speaker}) \tag{1}$$

The second analysis compared lexical /ɐ/+nasal to *um* (*uh* and lexical /ɐː/ were not examined here due to the larger number of *um* tokens and, as shown below, in the first analysis *um* was found to be significantly more likely to be creaky than *uh*). The initial model was the same as that in the previous analysis except word type (*um* versus lexical /ɐ/+nasal) was included as a fixed effect instead of word and an additional random intercept for word was included. In this analysis, tokens in the 'isolated' condition were removed due to low numbers of lexical items in this position (n=11). The final model syntax is shown in 2. 2732 tokens (*um*=1438; lexical=1294) were included.

$$\begin{aligned} \text{Creak} \sim \text{word type} * \text{LOTE-BG} + \text{word type} * \text{pos} + \\ \text{LOTE-BG} * \text{gender} + \text{pos} * \text{dur} + \\ (1 \mid \text{word}) + (1 + \text{word type} \mid \text{speaker}) \end{aligned} \tag{2}$$

3. Results

3.1. *Um* versus *Uh*

The output of model 1 is shown in Table 1. The model shows a significant effect of word, indicating that *um* tokens are more likely to be creaky than *uh* tokens. There was also a significant simple effect of duration showing that as tokens increased in duration, so did the likelihood of creak presence. Finally, there was a significant effect of position on the likelihood of creak being present, which is illustrated in Figure 1. Comparing the different position levels, tokens in phrase-initial position were significantly less likely to contain creak compared to all other positions. There were no significant differences in creak probability between tokens in isolated, phrase-medial and phrase-final positions; however, there was a non-significant trend towards tokens in medial position being more likely to contain creak compared to isolated (p=0.091) and final (p=0.063) positions.

Table 1: *GLMER model output showing effect of word, position in utterance and duration on creak presence. Significant effects are shaded. Ref. levels: word = uh; position = isolated.*

	Est.	Std. Err.	z-value	p-value
(Int.)	-0.45	0.16	-2.80	0.005
wordUm	0.43	0.14	3.07	0.002
posMed	0.25	0.15	1.69	0.091
posFin	-0.03	0.16	-0.16	0.876
posInt	-0.36	0.13	-2.77	0.006
dur	0.17	0.05	3.21	0.001

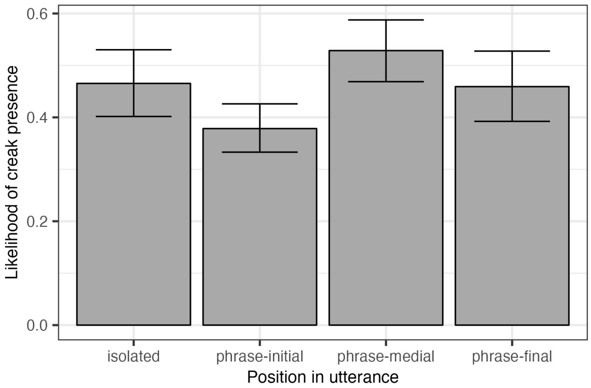


Figure 1: *GLMER model estimates of creak presence by position in utterance.*

3.2. *Um* versus lexical items

The output of model 2 is shown in Table 2. It shows a significant interaction between duration and position on creak likelihood. Figure 2 shows that tokens across all positions are generally more likely to be creaky the longer their duration; however, there is less of an effect of duration in initial position.

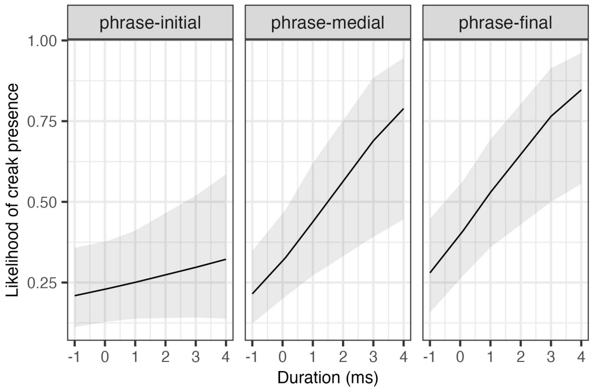


Figure 2: *GLMER model estimates of creak presence by position in utterance and token duration.*

Table 2: *GLMER model output showing effects of word type, LOTE-BG and position in utterance on creak presence. Significant effects are shaded. Ref. levels: word type = lexical; LOTE-BG = no; gender = female; position = phrase-medial.*

	Est.	Std. Err.	z-value	p-value
(Int.)	-0.38	0.39	-0.97	0.334
wordUm	0.54	0.68	0.79	0.428
LOTEYes	-1.68	0.41	-4.16	<0.001
genderM	-0.64	0.39	-1.64	0.101
posFin	1.16	0.27	4.27	<0.001
posInt	-0.48	0.46	-1.04	0.297
dur	0.52	0.17	3.07	0.002
wordUm:LOTEYes	1.30	0.37	3.52	<0.001
LOTEYes:genderM	1.10	0.44	2.48	0.013
wordUm:posFin	-1.53	0.42	-3.65	<0.001
wordUm:posInt	0.07	0.53	0.14	0.889
posFin:dur	0.01	0.24	0.04	0.973
posInt:dur	-0.41	0.20	-2.04	0.041

There was also a significant interaction of LOTE-BG and gender. *Post hoc* pairwise comparisons were carried out using the emmeans package [45] in R. These showed that regardless of word type, females with non-LOTE backgrounds were significantly more likely to creak than those with LOTE backgrounds (p<0.001). Among speakers with a LOTE background, males were more likely to have creak than females (p=0.033).

Additionally, there was a significant interaction between word type and LOTE-BG, which is illustrated in Figure 3. Pairwise comparisons show creak was significantly more likely in *um* compared to lexical items among speakers with a LOTE background (p=0.026), but no difference for non-LOTE background speakers. Within lexical items, creak was significantly more likely in speakers with a non-LOTE background (p=0.001); however, LOTE-BG did not significantly affect creak likelihood in *um*.

Finally, there was a significant interaction between word type and position in utterance on creak likelihood, shown in Figure 4. Pairwise comparisons indicate that lexical items in phrase-final position were more likely to be creaky than those in phrase-initial (p=0.001) or phrase-medial (p<0.001) position. Among *um* tokens, there were no significant differences.

4. Discussion

This study has explored whether creak presence differs between the fillers *um* and *uh*, and between *um* and lexical /ɐ/+nasal. Focusing first on the filler only analysis, no social variables affected

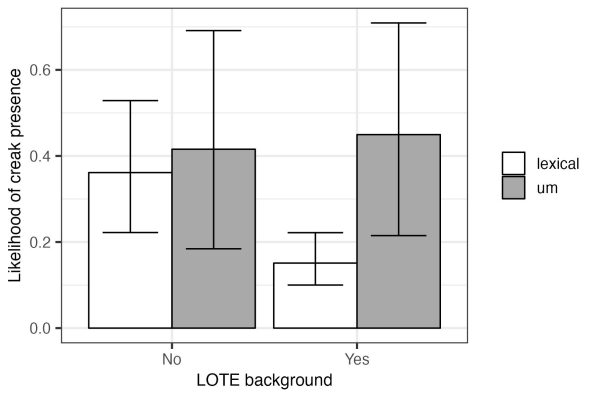


Figure 3: *GLMER model estimates of creak presence by word type and LOTE-BG.*

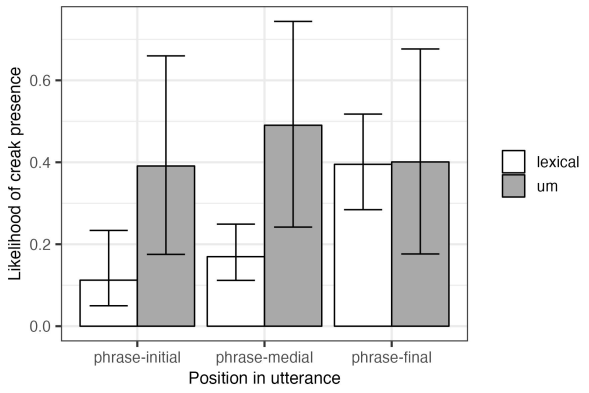


Figure 4: *GLMER model estimates of creak presence by word type and position in utterance.*

the use of creak suggesting that both male and female speakers from monolingual English and LOTE backgrounds use creak to similar degrees on *um* and *uh*. Tokens of *um* were significantly more likely to contain creak compared to *uh*. This supports findings from previous work showing that different fillers can vary in their acoustic characteristics, in this case voice quality [21, 11, 12]. The finding that both *um* and *uh* are less likely to contain creak in initial position is consistent with previous research showing that utterance-initial fillers have higher f0 than those later in the utterance [13, 21]. As low f0 is an acoustic characteristic of creak, it is possible that speakers are using creak to access their lower f0 range in tokens closer to the ends of utterances. Creak presence has been found to increase with phrase-finality (and turn-finality) across varieties of English, including AusE, and in other languages [27, 28, 29]. This could be a contributing factor to the duration effect where longer tokens were found to have higher likelihood of containing creak, due to the process of declination in f0: as airflow decreases across the course of an utterance (or in this case a long token), conditions that favour creak increase [26, 46]. It is important to bear in mind that this analysis does not take into account the location of the creak within a token and therefore further investigation of this point is needed. Although only approaching significance for isolated and final positions, the trend for medial tokens to have a higher likelihood of creak presence is consistent with research showing that clause-internal fillers are lower in f0 compared to their surrounding prosodic context [19, 20].

Turning to the analysis comparing *um* to lexical /ɐ/+nasal, the finding that female speakers with non-LOTE backgrounds were more likely to produce creaky tokens than female speakers with LOTE backgrounds, regardless of the type of item, is consistent with previous work on creak in AusE [31, 32]. While gender did not significantly impact creak by word type, LOTE-BG did have an effect. Interestingly, while speakers with a LOTE background had significantly less chance of creaking on lexical items compared to non-LOTE speakers, there was no difference in creak likelihood by LOTE-BG for *um*. This could suggest that on lexical items, creak can be used socially to signify aspects of a speakers' identity such as their cultural/linguistic heritage as suggested by [31]. However, creak is just as likely to occur on *um* for speakers of LOTE and non-LOTE backgrounds suggesting that these social meanings of creak are limited to when they occur on lexical items.

We now turn to significant interactions involving position in utterance. There was a significant interaction between position and duration regardless of word type. While tokens in initial position were less affected by duration when it comes to creak presence, those in medial or final positions were more likely to be creaky as duration increased. Again, this could be related to declination in f0: at the beginning of an utterance airflow is generally high whether tokens are short or long; however, as the utterance progresses, airflow decreases, increasing likelihood of creak, especially when tokens are long or drawn out [26, 46]. The interaction between word type and position suggests that while position has an effect on creak likelihood in lexical items, this is not the case for *um*. Higher probability of creak presence on lexical items in final position compared to initial or medial is consistent with previous studies suggesting creak is a marker of phrase-finality [27, 28, 29].

5. Conclusions

The present study has suggested that, in AusE, there is a relationship between creaky voice presence and the fillers *um* and *uh*. While previous work has shown that *um* and *uh* differ in their acoustics with regard to f0 and duration [11, 12, 21], we have shown that this finding extends to creaky voice quality. Additionally, our findings present evidence that *um* differs in the presence of creaky voice compared to lexical /ɐ/+nasal sequences. These results have potential implications for improving ASR technology if creaky voice quality can be used to aid in distinguishing between fillers and lexical items. It is also possible that creak could be considered alongside formant, duration and f0 information in using fillers for forensic use, especially considering recent research suggests interspeaker variation exists in the acoustics of creak in Dutch [47]. Results suggest that although macro-social factors play a role in the likelihood of creak occurring on lexical items, they do not have the same effect for *um*. This could be due to potential sociolinguistic functions of fillers such as conveying uncertainty [3, 4], which has also been linked to creaky voice [48]; however further research is required to explore this.

6. Acknowledgements

This research was supported by Australian Research Council Grant DP190102164 and Australian Research Council Future Fellowship Grant FT180100462 to the fifth author.

7. References

[1] Clark, H. H. and Fox Tree, J. E., "Using uh and um in spontaneous speaking," *Cogn.*, vol. 84, pp. 73–111, 2002.

[2] Levelt, W. J. M., "Monitoring and self-repair in speech," *Cogn.*, vol. 14, no. 1, pp. 41–104, 1983.

[3] Smith, V. L. and Clark, H. H., "On the course of answering questions," *J. Memory Lang.*, vol. 32, no. 1, pp. 25–38, 1993.

[4] Kirkland, A., Lameris, H., Székely, É., and Gustafson, J., "Where's the uh, hesitation? The interplay between filled pause location, speech rate and fundamental frequency in perception of confidence," in *Proc. Interspeech*, 2022, pp. 4990–4994.

[5] Tottie, G., "On the use of *uh* and *um* in American English," *Functions of Lang.*, vol. 21, no. 1, pp. 6–29, 2014.

[6] Foulkes, P., Carrol, G., and Hughes, S., "*Sociolinguistics and acoustic variability in filled pauses* [Conference presentation]," Ann. Conf. Int. Assoc. Forensic Phonetics and Acoustics, 2004.

[7] Tottie, G., "*Uh* and *Um* as sociolinguistic markers in British English," *Int. J. Corpus Linguistics*, vol. 16, no. 2, pp. 173–197, 2011.

[8] Acton, E. K., "On gender differences in the distribution of um and uh," *University of Pennsylvania Working Papers in Linguistics*, vol. 17, no. 2, p. article 2, 2011.

[9] Shriberg, E., "To 'errrr' is human: Ecology and acoustics of speech disfluencies," *J. Int. Phonetic Assoc.*, vol. 31, no. 1, pp. 153–169, 2001.

[10] Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., and Liberman, M., "Variation and change in the use of hesitation markers in Germanic languages," *Lang. Dyn. Change*, vol. 6, no. 2, pp. 199–234, 2016.

[11] Hughes, V., Wood, S. and Foulkes, P., "Strength of forensic voice comparison evidence from the acoustics of filled pauses," *Int. J. Speech, Lang. Law*, vol. 23, no. 1, pp. 99–132, 2016.

[12] Wong, S. G.-J. and Papp, V., "*Transferability of non-lexical hesitation markers across languages: Evidence from te reo Māori-English bilinguals* [Conference presentation]," 27th Ann. Conf. Int. Assoc. for Forensic Phonetics and Acoustics, 2018.

[13] de Boer, M. M. and Heeren, W. F. L., "Cross-linguistic filled pause realization: The acoustics of uh and um in native Dutch and non-native English," *J. Acoust. Soc. Am.*, vol. 148, pp. 3612–3622, 2020.

[14] Candea, M., Vasilescu, I., and Adda-Decker, M., "Inter- and intra-language acoustic analysis of autonomous fillers," in *Proc. DiSS*, 2005, pp. 47–52.

[15] Adell, J., Bonafonte, A. and Escudero-Mancebo, D., "Modelling filled pauses prosody to synthesise disfluent speech," in *Proc. Speech Prosody*, 2010.

[16] Belz, M. and Reichel, U. D., "Pitch characteristics of filled pauses," in *Proc. DiSS, ICPhS Satellite Meeting*, 2015.

[17] Maekawa, K., "Prediction of f0 height of filled pauses in spontaneous Japanese: a preliminary report," in *Proc. DiSS*, 2013, pp. 41–44.

[18] O'Shaughnessy, D., "Recognition of hesitations in spontaneous speech," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*. IEEE, 1992, pp. 521–524.

[19] Shriberg, E. E. and Lickley, R. J., "The relationship of filled-pause f0 to prosodic context," in *Proc. IRCS Workshop on Prosody in Natural Speech*, 1992, pp. 201–209.

[20] ——, "Intonation of clause-internal filled pauses," *Phonetica*, vol. 50, pp. 172–179, 1993.

[21] Swerts, M., "Filled pauses as markers of discourse structure," *J. Pragmatics*, vol. 30, pp. 485–496, 1998.

[22] Maekawa, K. and Mori, H., "Voice-quality difference between the vowels in filled pauses and ordinary lexical items," in *Proc. Interspeech*, 2016, pp. 3171–3175.

[23] Li, X., Ishi, C. T., Fu, C., and Hayashi, R., "Prosodic voice quality analyses of filled pauses in Japanses spontaneous conversation by Chinese learners and Japanese native speakers," in *Proc. Speech Prosody*, 2022, pp. 550–554.

[24] Maekawa, K., 'ya Nishikawa, K., and Tseng, S.-C., "Phonetic characteristics of filled pauses: A preliminary comparison between Japanese and Chinese," in *Proc. DiSS*, 2017, pp. 41–44.

[25] Keating, P., Garellek, M., and Kreiman, J., "Acoustic properties of different kinds of creaky voice," in *Proc. 18th ICPhS*, The Scottish Consortium for ICPhS 2015, Ed. University of Glasgow, 2015, Conference Proceedings, pp. 821.1–5.

[26] Podesva, R. J., "Gender and the social meaning of non-modal phonation types," *Ann. Meet. Berkeley Linguistics Soc.*, vol. 37, no. 1, pp. 427–448, 2013.

[27] Redi, L. and Shattuck-Hufnagel, S., "Variation in the realization of glottalization in normal speakers," *J. Phonetics*, vol. 29, no. 4, pp. 407–429, 2001.

[28] Ogden, R., "Turn transition, creak and glottal stop in finnish talk-in-interaction," *J. Int. Phonetic Assoc.*, vol. 31, no. 1, pp. 139–152, 2001.

[29] White, H., Penney, J., Gibson, A., Szakay, A., and Cox, F., "Creak prevalence and prosodic context in Australian English," in *Proc. Interspeech*, 2023.

[30] Loakes, D. and Gregory, A., "Voice quality in Australian English," *JASA Express Letters*, vol. 2, no. 8, p. 085201, 2022.

[31] White, H., Gibson, A., Penney, J., Szakay, A., and Cox, F., "*Differences in prevalence of creaky voice in mono- and multilingual communities in Sydney* [Conference presentation]," 14th Int. Symposium on Bilingualism, Sydney, Australia, 2023.

[32] White, H., "Creaky voice in Australian English," [PhD thesis, Macquarie University], 2023.

[33] Cox, F. and Penney, J., "Multicultural Australian English - The new voice of Sydney," *Aus. J. Linguistics*, forthcoming.

[34] White, H., Penney, J., Gibson, A., Szakay, A., and Cox, F., "Evaluating automatic creaky voice detection methods," *J. Acoust. Soc. Am.*, vol. 152, no. 3, pp. 1476–1486, 2022.

[35] Dorreen, K., "Fundamental frequency distributions of bilingual speakers in forensic speaker comparison," [Master's thesis, University of Canterbury], 2017.

[36] Dallaston, K. and Docherty, G., "Estimating the prevalence of creaky voice: A fundamental frequency-based approach," in *Proc. 19th ICPhS*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. ASSTA Inc, 2019, pp. 581.1–5.

[37] Drugman, T., Kane, J., and Gobl, C., "Data-driven detection and analysis of the patterns of creaky voice," *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1233–1253, 2014.

[38] Schiel, F., Draxler, C., and Harrington, J., "Phonemic segmentation and labelling using MAUS technique," in *New tools and methods for very-large-scale phonetics research workshop*, Philadelphia, PA, USA, 2011.

[39] Boersma, P. and Weenink, D., "Praat: Doing phonetics by computer [Computer program]," 2023, version 6.3.09. Online: http://www.praat.org/

[40] Garellek, M., "Perception of glottalization and phrase-final creak," *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. 822–831, 2015.

[41] Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M., "Glottalization of word-initial vowels as a function of prosodic structure," *J. Phonetics*, vol. 24, no. 4, pp. 423–444, 1996.

[42] Bates, D., Maechler, M., Bolker, B., and Walker, S., "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[43] Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B., "lmerTest package: Tests in linear mixed effects models," *J. Stat. Softw.*, vol. 82, no. 13, pp. 1–26, 2017.

[44] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2022. Online: https://www.R-project.org/

[45] Lenth, R., "emmeans: Estimated marginal means, aka least-squares means," 2018, R package version 1.8.9. Online: https://CRAN.R-project.org/package=emmeans

[46] Ladd, D. R., "Declination: A review and some hypotheses," *Phonology Yearbook*, vol. 1, pp. 53–74, 1984.

[47] van Hugte, T. B. R. and Heeren, W. F. L., "Exploring interspeaker variation in creaky voice in Dutch," *J. Voice*, 2024.

[48] Yuasa, I. P., "Creaky Voice: A New Feminine Voice Quality for Young Urban-Oriented Upwardly Mobile American Women?" *Am. Speech*, vol. 85, no. 3, pp. 315–337, 2010.

The Effects of Syntactic Dependencies and Speech Tempo on Macro-Rhythm

Catalina Torres[1], *Marie-Anne Morand*[2], *Sebastian Sauppe*[1], *Balthasar Bickel*[1]

[1]University of Zurich, [2]University of Fribourg

catalina.torres@ivs.uzh.ch, marie-anne.morand@unifr.ch,
sebastian.sauppe@psychologie.uzh.ch, balthasar.bickel@uzh.ch

Abstract

Macro-rhythm is a parameter that distinguishes the prosodic profiles of languages at the intonational level. However, no conclusive quantification of macro-rhythm differences between languages has been proposed and there are to date no dedicated studies investigating the regularity and variability of macro-rhythm within a single language. This study experimentally examines the global tonal patterns of Swiss German to test (i) internal effects related to word order and (ii) articulation rate effects on global tonal patterns. We find that macro-rhythm is affected by language-internal syntactic variation and by intra-speaker variability (in terms of speech tempo).

Index Terms: macro-rhythm, prosody, word order, verb clusters, crossed/nested/adjoined dependencies, Swiss German

1. Introduction

This study analyses the relationship between intonational phonology and syntactic dependencies by examining global tonal patterns within the autosegmental-metrical (AM) approach [1, 2, 3]. The goal is to investigate the correspondence between intonational phonology and syntactic linearisation in Zurich German, a variety of German spoken in Switzerland allowing for variable word order in complex verb clusters. In AM, tonal space refers to an interval where the high (H) tones are at the top level and the low (L) tones are at the bottom level of a band of fundamental frequency (F0) values [2]. Variations in tonal space, such as intonational boundary tones, can serve as a phonological representation of prosodic constituency [4] or interact with other linguistic domains, such as syntactic structure. For example, a major syntactic break can be associated with an intermediate phrase boundary [5, 6]. In addition to signaling prosodic prominence and boundary marking, prosodic macro-rhythm [3] has also been proposed as a third parameter distinguishing the prosodic profiles of languages. Macro-rhythm is the temporal organisation of speech, perceived by the regular occurrence of tonal events in the F0 contour (i.e., rising and falling tonal movements, see Figure 1). At the intonational level, macro-rhythm differs in three dimensions [3]: (i) the *number of possible accents*, with larger accent inventories predicting more tonal variability and therefore lower degrees of macro-rhythm; (ii) the *type of most common accents*, with rising or falling accents producing higher degrees of macro-rhythm than level accents; and (iii) the *frequency or domain of accents*, with more accents in a phrase resulting in a higher degree of macro-rhythm, because every smaller domain (e.g., every prosodic word) receives an accent.

Cross-linguistic studies on macro-rhythm have focused on F0 measures taken from speech corpora and used different calculations, with mixed results [7, 8, 9]. While nPVI calculates the variability in the distance of intervals between F0 peaks and valleys, MacR_Var focuses on the standard deviations of rising or falling slopes, peak-to-peak distance and valley-to-valley distance per Intonation Phrase. However, these measures do not account for possible variation arising from the use of different syntactic structures within a language or from intra-speaker variability in the rate of articulation. The effects arising from the number of possible accents may be less important in predicting macro-rhythm compared to the type of the most common accent and their frequency domain [8]. Here, we examine the influence of syntax on the number of possible accents per phrase, while also considering the most common accents in Swiss German. Although boundary synchronisation between prosodic and syntactic constituents has been documented, little is known about how syntax interacts with macro-rhythm and phrase-level global tonal patterns. Rising and falling tonal movements play an important role in attention orienting. For example, neurophysiological studies have shown that rises in amplitude of pure sine tones are linked to auditory looming effects [10] and that tonal rises, in particular, guide auditory attention when listening to speech [11]. We examine the relationship between verb clusters and macro-rhythm testing whether word order has an effect on intonational patterning.

1.1. Zurich German

In German-speaking Switzerland, people usually use at least two language varieties of German regularly: (Swiss) Standard German and a Swiss German dialect. In contrast to other German-speaking countries, there is no standard-dialect continuum [12]. The sociolinguistic situation can thus be described as *diglossia* [13, 14, 15] with rather clear-cut boundaries between the standard and the vernacular varieties. Previous studies of prosodic features in Swiss German focused on temporal aspects, such as speech rhythm and speech tempo [16, 17, 18], and little is known about intonational patterns. Compared to (northern) Standard German, Zurich German shows a larger overall F0 range, with a greater number of pitch movements and a default pitch accent (L*+H) consisting of a low–rising contour, and a slower speech rate [19]. Thus, Zurich German can be characterised as having a stronger macro-rhythm, compared to Standard German with its medium degree of macro-rhythm [3].

1.2. Complex verb clusters in Zurich German

This study examines adjacent and non-adjacent syntactic dependencies (i.e., relations between sentence elements) and how they are prosodically marked. Dependencies can be expressed in different ways: Either so that dependency arcs cross each other (as in the pattern *ABAB*, where the *A*s and the *B*s are non-adjacent, cf. Figure 2-1) or so that the elements and their dependencies are nested within each other (as in *ABBA*, where the *B*s are adjacent, but not the *A*s, cf. Figure 2-2) [20]. A third possible realisation are adjoined dependencies (as in *AABB*, where each dependency arc is immediately closed, cf. Figure 2-3).

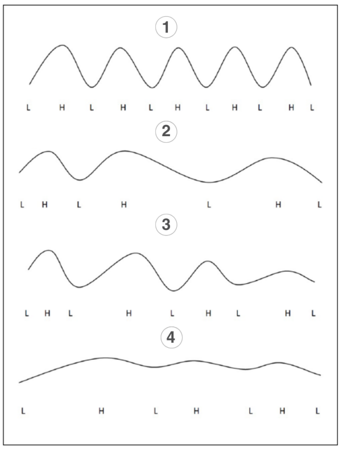
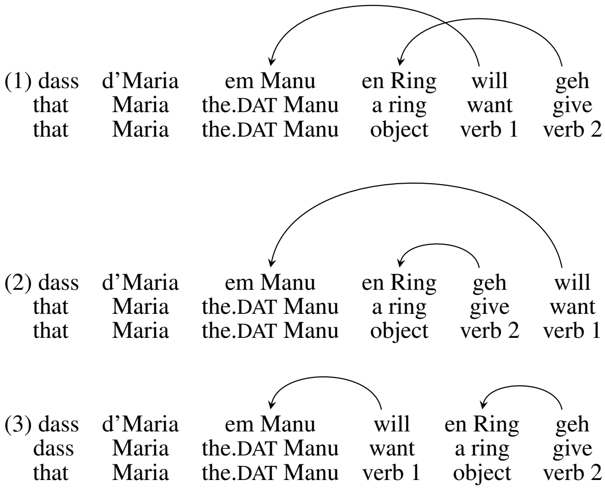


Figure 1: Schematic pitch contours showing differences in the temporal (1–2) and frequency domain (3–4) of macro-rhythm (adapted from [7]). Contour 1 exhibits H and L turning points that are more evenly distributed in time and thus has a stronger macro-rhythm than contour 2. Contour 3 exhibits larger differences in frequency between successive H and L turning points and thus creates a stronger macro-rhythm than contour 4.

Verb clusters (or verbal complexes) are verb phrases that contain several verbs. In German, these verbs are usually a finite auxiliary or modal verb (*will* 'want' in 2), a main verb in the infinitive form (*geh* 'give' in Figure 2), and its object if it is transitive (*en Ring* 'a ring' in Figure 2). While any lexical verb can function as the infinitive main verb, there is only a small, closed class of auxiliaries and modals that can function as finite verb in verbal complexes. Non-adjacent dependencies bear great relevance in linguistics because they have played a major role in Formal Language Theory (FLT), which describes which computational properties grammars of human languages need to have [21, 22]. In the Chomsky hierarchy in FLT, syntactic structures with crossed dependencies require more complex computations to be describable, thus requiring context-sensitive grammars [21]. Crossed dependencies are cross-linguistically rare and have so far only been attested in Dutch and Swiss German [23, 24]. We take advantage of the flexibility of word order in the Swiss German verbal complex to explore whether there are prosodic correlates of syntactic structure that signal which type of dependency a currently uttered sentence has. This is possible because nested, crossed, or adjoined dependencies can be used to express the same meaning (Figure 2). Signaling the difference between crossed and nested dependencies could, for example, be helpful to listeners, who are known to form expectations about the upcoming linguistic input [25, 26] and thus would potentially benefit from cues about the dependency type.

1.3. Research questions

The variable word order in Swiss German verb clusters offers the possibility to test how intonational structure is realised in relation to syntax, while keeping the semantic information constant. Considering the reported relevance of rising tonal movements for guiding perception [11], it is conceivable that the intonational structure of complex syntactic structures is organised to facilitate their parsing. In this study, we therefore ask the following questions:



(1) dass	d'Maria	em Manu	en Ring	will	geh
that	Maria	the.DAT Manu	a ring	want	give
that	Maria	the.DAT Manu	object	verb 1	verb 2

(2) dass	d'Maria	em Manu	en Ring	geh	will
that	Maria	the.DAT Manu	a ring	give	want
that	Maria	the.DAT Manu	object	verb 2	verb 1

(3) dass	d'Maria	em Manu	will	en Ring	geh
dass	Maria	the.DAT Manu	want	a ring	give
that	Maria	the.DAT Manu	verb 1	object	verb 2

Figure 2: Three verb cluster orders and their dependency structure in Zurich German for the clause '…that Maria wants to give Manu a ring'. (1) crossed, (2) nested, (3) adjoined.

R1 What are the global tonal patterns of utterances with the three different word orders in the verb cluster (crossed, nested, adjoined)?

R2 Do these global tonal patterns differ?

R3 If global tonal patterns differ, how do they differ?

2. Materials and Methods

Stimulus sentences were created by a native speaker of Swiss German, specialist of Zurich German. Subsequently, these sentences were checked for comprehensibility by a native speaker of Zurich German. Sentences considered difficult to comprehend were discarded before data collection. To create stimuli that allowed the recognition of intonation contours, the sentences contained as many sonorant consonants as possible. A set of 55 'dass' frame sentences were created and the three possible verb clusters were used (cf. Figure 2). To test whether sentence length has an effect on the intonational contour we created stimuli of varying length. The sentences vary in the number of syllables they contain, in a scale of five steps each sentence becomes longer by one additional syllable. This count only includes the syllables of the words that can be accented. This means, that in the region of interest, we have contours with 5 to 10 syllables. As there is no standard orthography for Zurich (or Swiss) German, our stimuli were made using a writing system that could be used, for instance, when chatting on social media and which had been tested in another study before.

2.1. Participants

Ten speakers (5 female) of Zurich German were invited to the study. Speakers were 20-33 years old (mean = 24.2 years) and reported growing up in the canton of Zurich, as well as living in the city of Zurich (or near by) at the time of recording.

2.2. Recordings

Participants were recorded at a self-selected normal and fast speech rate and compensated with study credit. Before recording, the participants obtained a printed copy containing all the sentences so that they could become familiar with them and ask questions if anything was unclear. They were seated in front of a screen in a sound-attenuated booth and were prompted with a written sentence they had to read out loud, using ProRec software (Mark Huckvale, University College London). If they felt they were not fluent or misread the sentence, the recording was repeated. The recordings were made at a sampling rate of 44.1 kHz and 16 bit, using a Røde 1000 large diaphrame condensor microphone.

2.3. Analysis procedures

To obtain a transcription at the utterance level, a bash script was used. Text files and associated WAV files were used for forced alignment of the speech signal via the web interface of the Munich Automatic Segmentation System using the CH (German Dieth) language model. For forced alignment, the web service G2P was used to convert the orthographic text input into a canonical phonological transcript [27]. The resulting files were used to compute a phonetic segmentation and labelling based on the speech signal and a phonological transcript in webMAUS [28]. The alignment process provided a TextGrid for each utterance in which all words were segmented and marked with boundaries. Data was further processed using Praat 6.3.06 [29]. There were a total of 1556 utterances in the corpus[1].

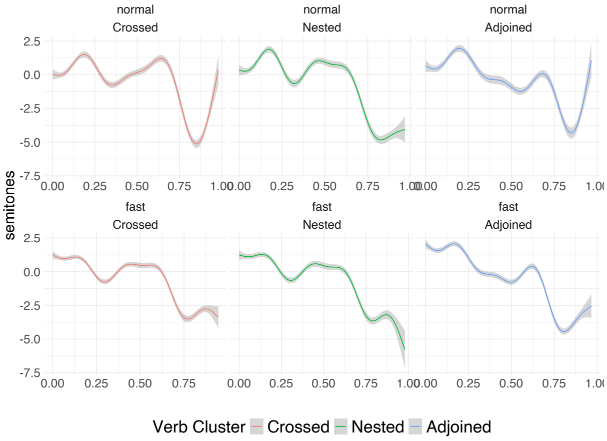


Figure 3: Smoothed fundamental frequency trajectories of three verb clusters in semitones of normal (top) versus fast (bottom) speech rate. The start of the word *Manu* represents the onset.

We are interested in the intonational contours at the phrasal level. Figure 3 shows F0 contours for the three verb clusters. To obtain the F0 contours, 150 consecutive measurements were taken per utterance, starting with the onset of the word *Manu* (cf. Figure 2), up to the end of the utterance. As described in §2, we created our speech materials to allow for a continuous F0 contour with minimal devoiced segmental material (which could not be entirely avoided). However, there are unavoidable phonetic perturbations that are not intended when the F0 movements are planned during speech production and which can lead to inaccurate F0 measurements. As the F0 contour can be affected by short-term perturbations caused by segmental characteristics such as junctures between consonants and vowels, variations in voice quality, or F0 tracking errors, we decided to use an automated method for detecting F0 measurement jumps based on sample-to-sample differences [30]. Utterances produced with intervening pauses were discarded to avoid effects resulting from pitch reset in relation to prosodic boundaries. To test whether forced alignment provided a reliable segmentation of words, a subset of the data was hand corrected (31% of the corpus). The onset and offset time stamps of the word *Manu* were queried and statistically evaluated using Wilcoxon rank sum tests. The mean onset in forced-aligned words (840 ms) was not significantly different from that of the hand-corrected sample (836 ms) ($p = 0.6$, effect size $r = 0.015$). A similar result was obtained for the offset ($p = 0.4$, effect size $r = 0.026$), confirming the reliability of forced-aligned utterances.

2.3.1. Statistical analysis

The F0 trajectories are analysed with Generalised Additive Mixed Models (GAMMs) in R [31], using the `mgcv` [32]. We include syntax and sex as parametric predictors and additional random effects for speaker and sentence ID. Following [33], we created new (factor smooth) variables, representing the interaction between verb cluster type and the number of syllables per clause and representing the interaction between verb cluster type and speech rate.

3. Results

The results of the GAMMs show that the tonal global patterns systematically differ between the three verb clusters.

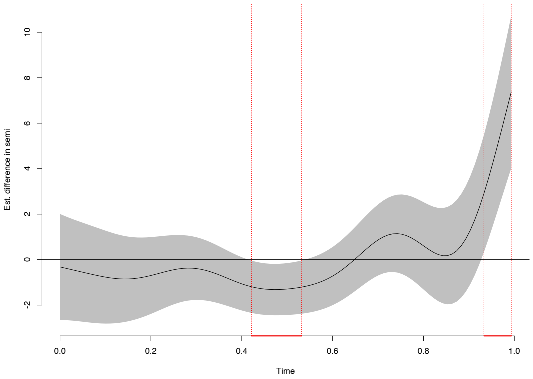


Figure 4: Difference smooths contrasting crossed versus nested verb clusters in short utterances. The pointwise 95%-confidence interval is shown by the shaded band.

Additionally, we find a statistically significant difference for the manipulation of speech rate. As shown in Figure 3, contours at a normal speech rate show more expanded tonal movements (i.e., stronger macro-rhythm), while the F0 range is narrower at faster speech rates (i.e., weaker macro-rhythm). Figures 4–6 show the fitted difference smooths between verb cluster types. When the estimated difference is significantly different from zero (i.e., when the 95% confidence interval for the difference between smooths does not include 0), this is indicated by a red line on the x-axis and vertical dotted lines on the y-axis. First, the onset word *Manu* (held constant across all stimuli) does not show any significant difference. However, the tonal patterns differ significantly towards the middle and end portions of the contours. Figure 4 shows the estimated difference between short crossed and nested verb clusters. This illus-

[1] For two female participants recorded in a pilot there were only 48 frame sentences (288 in total).

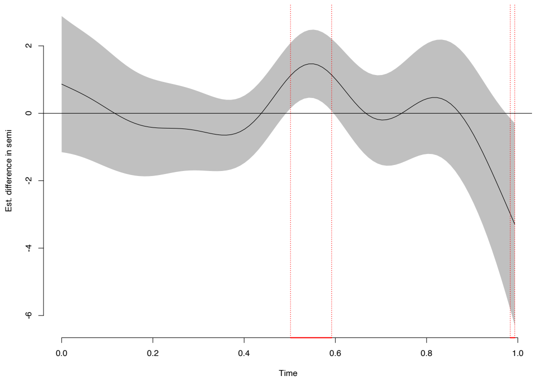


Figure 5: Difference smooths contrasting crossed versus adjoined verb clusters in short utterances. The pointwise 95%-confidence interval is shown by the shaded band.

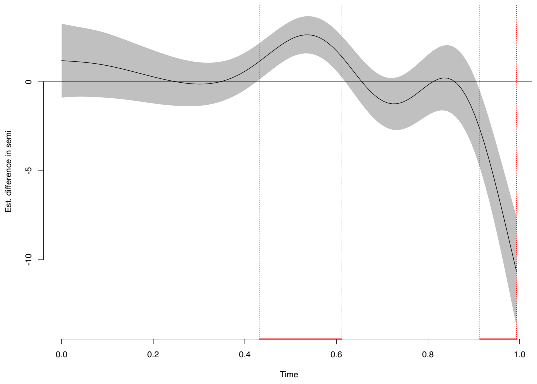


Figure 6: Difference smooths contrasting nested versus adjoined verb clusters in short utterances. The pointwise 95%-confidence interval is shown by the shaded band.

trates a significant difference towards the midpoint and the end of the utterances, whereby the contour in the crossed clusters is lower in the middle but higher at the end compared to the nested condition. Figure 5 shows the estimated difference between the crossed and adjoined verb clusters whereby the intonational contour of crossed verb clusters is higher at the midpoint and towards the end compared to nested verb clusters. Figure 6 shows the estimated difference between the adjoined and nested verb clusters. In this case, the intonation contour in the nested condition appears to be higher at midpoint but lower towards the end compared to the adjoined contours. Taken together, the contours vary due to word order, despite all sentences sharing the same lexical material. This is evident in the short utterances in which all content words in the verb cluster are monosyllabic (see Figure 2). Additionally, the differences in global tonal pattern are mostly robust for further comparisons between crossed versus nested and nested versus adjoined. However, in longer utterances comparisons between crossed and adjoined are not significantly different.

4. Discussion and Conclusion

Although macro-rhythm has been proposed as an additional parameter to distinguish languages prosodically [3], it has proved to be difficult to measure in acoustic terms. We investigated two factors that may have an influence on macro-rhythm within a single language, focusing on the role of language-internal word order alternations and on speech tempo variation. We find that an interaction between syntax and prosody modulates macro-rhythm and that speech tempo affects measures of macro-rhythm through the magnitude of tonal movements. It is likely that, similar to rhythm metrics [34], the (ir-)regularity of macro-rhythm is related to perceptual rather than acoustic factors. We note that in our data, global tonal effects seem to be driven by the predictability of lexical material, whereby less predictable content words are realised with rising tonal movements and more predictable verbs are deaccented. In a comparison of the intonational contours of three different verb clusters in Zurich German (crossed, nested, and adjoined), we find differences between all of them. As shown in Figure 3, within the portion of the dependencies (starting with *Manu*, Figure 2), we find three tonal rises in the crossed and adjoined contours at a global level. In comparison, the nested verb clusters show only two tonal rises. Additionally, the nested contours in Figure 3 also show deaccentuation of the sentence-final finite verb (*will* 'want', Figure 2). Although the current methodology does not allow us to reconstruct the exact position of segments, the controlled nature of the stimulus material makes it possible to estimate the most important points in the utterance, where we observe significant variation.

Out of the three verb clusters studied, the nested structure shows a lower macro-rhythm. Important differences arise in the region of the verb clusters (excluding the tonal rise in Manu, see Figure 2). We find that intonational differences are modulated by the height of the contour (where we predict L*+H pitch accents) on the object and the infinite verb 2 (*geh* 'give', Figure 2) while the finite verb (*will* 'want', Figure 2) shows deaccentuation. One possible explanation for these differences is that the number of verbs that can occur as the finite verb in complex verbal clusters is limited and therefore more predictable. In contrast, the object and infinite verb in these constructions can be freely chosen and are thus less predictable. If we consider the relevance of rising tonal movements in guiding auditory perception when listening to speech material [11], this can explain why the less predictable speech material is accented (object, infinite verb) while the more predictable speech material (restricted set of possible finite verbs) is not. In the related variety of Bernese German [35], deaccentuation was reported for words out of focus. This is in line with our observation that the more informative material is realised with a rising intonation whereas the less informative material is deaccented. The experimental investigation shows the effects of syntax and speech tempo on global tonal patterns. Our results show that macro-rhythm is sensitive to language internal structure provided by word order, whereby some structures can lead to an increased number of tonal movements. Additionally, we show that variations of speech tempo have an effect on the contour, whereby a fast speech tempo leads to a narrower F0 range. These results can help us understand why previous measures led to mixed results. First, we find that variable word order of otherwise equal lexical material leads to variation in global tonal patterns. Thus, not only the number of words per utterance but also how these words are syntactically structured plays an important role. Second, although intra-speaker variation has been acknowledged as a possible confound [9, 8], its role had not been demonstrated before. Taken together, we show that variable word order in verb clusters influences global tonal patterns and that speech tempo modulates pitch range in Zurich German.

5. Acknowledgements

We thank Sabrina Ryffel and Giorgia Sironi for help during data collection and Pelin Durtschi for her support in hand correcting segmentation, as well as Chantal Oderbolz for support with the creation of stimuli. We also thank the Linguistic Research Infrastructure (LiRI) at the University of Zurich for technical support. This work was funded by the University of Zurich and the Swiss National Center for Competence in Research "Evolving Language" (Swiss National Science Foundation agreement number 51NF40_180888, B.B.). S.S. is supported by the Swiss National Science Foundation Grant PZ00P1_208915.

6. References

[1] J. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, 1980.

[2] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[3] S.-A. Jun, *Prosodic typology: by prominence type, word prosody, and macro-rhythm*. Oxford University Press, 2014, pp. 520–540.

[4] H. Truckenbrodt, F. Sandalo, and B. Abaurre, "Elements of Brazilian Portuguese intonation," *Journal of Portuguese Linguistics*, vol. 8, no. 1, pp. 75–114, 2009.

[5] M. D'Imperio and A. Michelas, "Pitch scaling and the internal structuring of the Intonation Phrase in French," *Phonology*, vol. 31, no. 1, pp. 95–122, 2014.

[6] C. Torres, J. Fletcher, and G. Wigglesworth, "Fundamental frequency and regional variation in Lifou French," *Language and Speech*, vol. 65, no. 4, pp. 889–922, 2022.

[7] L. Polyanskaya, M. G. Busà, and M. Ordin, "Capturing cross-linguistic differences in macro-rhythm: The case of Italian and English," *Language and Speech*, vol. 63, no. 2, pp. 242–263, 2020.

[8] C. Kaland, "Bending the string: intonation contour length as a correlate of macro-rhythm." in *Interspeech*, 2022, pp. 5233–5237.

[9] C. Prechtel, "Macro-rhythm in English and Spanish: Evidence from radio newscaster speech," *Speech Prosody 2020*, pp. 675–679, 2020.

[10] D. R. Bach, H. Schächinger, J. G. Neuhoff, F. Esposito, F. D. Salle, C. Lehmann, M. Herdener, K. Scheffler, and E. Seifritz, "Rising sound intensity: an intrinsic warning cue activating the amygdala," *Cerebral Cortex*, vol. 18, no. 1, pp. 145–150, 2008.

[11] M. Lialiou, M. Grice, C. T. Röhr, and P. B. Schumacher, "Auditory processing of intonational rises and falls in German: rises are special in attention orienting," *Journal of cognitive neuroscience*, vol. 36, no. 6, pp. 1099–1122, 2024.

[12] U. Ammon, "Dialektschwund, Dialekt-Standard-Kontinuum, Diglossie: Drei Typen des Verhältnisses Dialekt-Standardvarietät im deutschen Sprachgebiet," in *Standardfragen: Soziolinguistische Perspektiven auf Geschichte, Sprachkontakt und Sprachvariation*, J. Androutsopoulos and E. Ziegler, Eds. Lang, 2003, pp. 163–171.

[13] C. A. Ferguson, "Diglossia," *word*, vol. 15, no. 2, pp. 325–340, 1959.

[14] J. A. Fishman, "Bilingualism with and without diglossia; diglossia with and without bilingualism," *Journal of Social Issues*, vol. 23, no. 2, pp. 29–38, 1967.

[15] G. Kolde, *Sprachkontakte in gemischtsprachigen Städten: vergleichende Untersuchungen über Voraussetzungen und Formen sprachlicher Interaktion verschiedensprachiger Jugendlicher in den Schweizer Städten Biel/Bienne und Fribourg/Freiburg i. Ue.* Steiner, 1981.

[16] A. Leemann, V. Dellwo, M.-J. Kolly, and S. Schmid, "Rhythmic variability in Swiss German dialects," in *Proceedings Speech Prosody 2012*, 2012, pp. 607–610.

[17] U. Zihlmann, "Vowel and consonant length in four Alemannic dialects and their influence on the respective varieties of Swiss Standard German," *Wiener Linguistische Gazette*, vol. 86, pp. 1–46, 2020.

[18] M.-A. Morand, M. Bruno, S. Schwab, and S. Schmid, "Syllable rate and speech rhythm in multiethnolectal Zurich German: A comparison of speaking styles," in *Proceedings Speech Prosody 2022*, 2022, pp. 337–341.

[19] J. Fleischer and S. Schmid, "Zurich German," *Journal of the International Phonetic Association*, vol. 36, no. 2, pp. 243–253, 2006.

[20] M. H. de Vries, K. M. Petersson, S. Geukes, P. Zwitserlood, and M. H. Christiansen, "Processing multiple non-adjacent dependencies: evidence from sequence learning," *Philosophical Transactions of The Royal Society B*, vol. 367, no. 1598, pp. 2065–2076, 2012.

[21] G. Jäger and J. Rogers, "Formal language theory: refining the Chomsky hierarchy," *Philosophical Transactions of The Royal Society B*, vol. 367, no. 1598, pp. 1956–1970, 2012.

[22] W. T. Fitch and A. D. Friederici, "Artificial grammar learning meets formal language theory: an overview," *Philosophical Transactions of The Royal Society B: Biological Sciences*, vol. 367, no. 1598, pp. 1933–1955, 2012.

[23] S. M. Shieber, "Evidence against the context-freeness of natural language," *Linguistics and Philosophy*, vol. 8, no. 3, pp. 333–345, 1985.

[24] J. Reese, *Swiss German: The Modern Alemannic Vernacular in and around Zurich*. München: Lincom Europa, 2007.

[25] G. R. Kuperberg and T. F. Jaeger, "What do we mean by prediction in language comprehension?" *Language, Cognition and Neuroscience*, vol. 31, no. 1, pp. 32–59, 2016.

[26] K. D. Federmeier, "Thinking ahead: The role and roots of prediction in language comprehension," *Psychophysiology*, vol. 44, no. 4, pp. 491–505, 2007.

[27] U. D. Reichel and T. Kisler, "Language-independent grapheme-phoneme conversion and word stress assignment as a web service," *Studientexte zur Sprachkommunikation*, pp. 42–49, 2014.

[28] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[29] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.3.06)[computer program]." 2023.

[30] J. Steffman and J. Cole, "An automated method for detecting F0 measurement jumps based on sample-to-sample differences," *JASA Express Letters*, vol. 2, no. 11, p. 115201, 11 2022. [Online]. Available: https://doi.org/10.1121/10.0015045

[31] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[32] S. N. Wood, "mgcv: GAMs and generalized ridge regression for R," *R news*, vol. 1, no. 2, pp. 20–25, 2001.

[33] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018.

[34] A. Arvaniti, "Rhythm, timing and the timing of rhythm," *Phonetica*, vol. 66, no. 1-2, pp. 46–63, 2009.

[35] J. Fitzpatrick-Cole, "The alpine intonation of Bern Swiss German," in *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 1, 1999, pp. 941–944.

Speech Task and Prosodic Context Influence Glottal Stop Variation in Tahitian

Janet Fletcher, Adele Gregory

School of Languages and Linguistics, University of Melbourne
janetf@unimelb.edu.au;adele.gregory@unimelb.edu.au

Abstract

It is well accepted that phonemic /ʔ/ can have multiple variants ranging from a canonical voiceless stop with full glottal occlusion to more vowel-like non-modal voiced variants. This study examines acoustic correlates of glottal stop variation in Tahitian. While it has a number of glottal variants like closely related Hawaiian, it differs significantly in that the major allophone in Tahitian is a fully or partially occluded glottal stop. While like Hawaiian more vowel-like variants are realised in less-constrained discourse, more voiceless variants and lower levels of Harmonics-to-Noise ratio are observed in word initial-versus word medial contexts.

Index Terms: glottal stop, Tahitian, duration, Harmonics-to-Noise ratio

1. Introduction

Tahitian (Reo Tahiti) is the dominant indigenous language of French Polynesia. It is part of the Eastern Polynesian sub-group of Polynesian languages [1]. It is estimated that between 60-70% of the population of French Polynesia speak, read, or write Tahitian or one of the other six indigenous languages of the archipelago. Tahitian is the most widespread in terms of daily usage although undergoing a process of reduced transmission from older to younger generations [2].

Table 1. *Phoneme inventory of Tahitian*

Consonants

	labial	*dentialveolar*	*Glottal*
Stops	p	t	ʔ
Nasals	m	n	
Fricatives	f v		h
Liquids		ɾ	

Vowels

	front		back
High	i iː		u uː (+Sequences)
Mid	e eː		o oː
Low		a aː	

The sound system of Tahitian is often described as "simple" (e.g. [3], [4], [5]) because of the small number of contrastive consonant and vowel phonemes relative to French or English, for example. Table 1 illustrates the consonant and vowel inventories of Tahitian (after [4]). The inventory consists of nine consonant phonemes with a simple stop series /p/ /t/ /ʔ/, five primary vowels that contrast in quantity and a number of vowel sequences/diphthongs. It is widely claimed that Oceanic languages like Tahitian have undergone a high degree of "phonetic erosion" over time (e.g. [6]). One key example that is often cited is the case of the glottal stop /ʔ/ which is an important contrastive phoneme in Tahitian and the other indigenous languages of French Polynesia. To date, there has been very limited experimental phonetic investigation of any aspect of spoken Tahitian and this paper focuses on variation in the acoustic realisation of the glottal stop /ʔ/.

It is commonly accepted that whilst complete closure of the glottis is the textbook realisation of a glottal stop, there can be variable realisations within and across the world's languages [7, 8]. Glottal stop variants range from a period of complete glottal occlusion to a period of creaky voice or even breathy voice along a continuum. It is also well attested that prosodic conditioning can also influence glottal stop realisation with more fully occluded or gesturally strengthened variants in Intonational Phrase-initial or word-initial positions [9,10]. In a large-scale survey of 131 Illustrations of the International Phonetic Alphabet published in the Journal of the International Phonetic Association, [8] showed that acoustic measures of glottal state including percentage of voicing and voicing intensity using strength of excitation measures (SoE) were effective in quantifying differences among glottal stop variants with more vowel-like variants showing higher % voicing and stronger levels of voicing compared to glottal stops. Findings were less clear for phrase or word position however. Similarly, in Arapaho, [11] found that harmonics-to-noise ratio (HNR) was a useful indicator of different glottal stop variants with more modal phonation showing a higher ratio compared to more creaky, laryngealised phonation. Fully occluded glottal stops show similar levels of HNR to voiceless /t/ in Arapaho and are always acoustically longer than other variants.

In Hawaiian, another Eastern Polynesian language, an investigation of glottal stop variation in naturalistic speech data showed that relatively few /ʔ/ are produced with full glottal closure (7%) with most variants produced with creaky voice variants[12]. These include a variant that consists of a period of creaky voice flanked by two modal vowels: 'mid-creak', and a variant 'whole creak' where an entire VʔV sequence is produced with creaky voice. Other more vowel-like realisations of /ʔ/ include a so-called 'intensity dip' variant that is a period of modal phonation of reduced intensity compared to surrounding modal vowels. Two further 'vocalic' variants noted by [12] are modal voice and breathy voiced variants. By contrast word-initial /ʔ/ is more likely to be produced with at least partial if not complete glottal closure suggesting a degree of prosodic strengthening in this location. Similar to Arapaho, however, full glottal closure variants tend to be somewhat longer than other more vowel-like variants in Hawaiian.

As yet, there has been no focused experimental study of glottal stop variation in Tahitian. One study [3] suggests that glottal stops are always voiceless but also mentions a variant that may well be similar to the so-called intensity dip noted by [12] for Hawaiian. [3] suggests that more detailed phonetic analyses are required. [13] also discusses laryngealised variants but concludes that variation is somewhat unpredictable whereas [4] claims that modal and more creaky voice realisations are likely when flanking vowels are identical in a VʔV sequence. In view of this, and also findings for the closely related language Hawaiian [12] and other languages that have a phonemic glottal stop [8][11], it is highly likely that Tahitian exhibits phonetic variation in /ʔ/ realisation. It remains to be seen whether prosodic context is a relevant factor in determining whether there are more voiceless, fully occluded variants in prosodic constituent-initial position, for example. If prosodically conditioned variation is present, we predict that more controlled laboratory phonology-type tasks where narrow focus is invariably realised on specific experimental tokens of interest may also result in more fully-occluded variants compared to connected speech [12]. In fact we would predict that the proportion of full closure glottal stops to more "vocalic" creaky voice variants in connected speech tasks will be similar to other studies that have focused on more naturalistic discourse.

2. Methodology

2.1. Speakers and materials

Speech data were obtained from five female speakers aged between 18 and 50 at the time of the recordings. All participants were born in French Polynesia and are L1 speakers of Tahitian along with Tahitian French. The recording materials consisted of two elicited reading tasks. The first set of materials (Experiment 1, henceforth EXP1) were designed to illustrate all contrastive stops and vowels, and consisted of two disyllabic words written in Tahitian orthography inserted in a carrier phrase. Ninety words were included in the study but only tokens containing the glottal stop or oral stops will be examined in this analysis. An example of the EXP1 carrier phrase is shown in (1). Tahitian orthography uses ' to indicate a glottal stop or *'eta*.

(1) I roto i te reo tahiti, e parauhia ***pata*** e'ere ***pa'a***.
"In Tahitian we say **scorpion** and not **bark**."

The second experimental task consisted of a reading and retelling of the Aesop fable La Bise et Le Soleil (the north wind and the sun), henceforth BES, following the materials used in the Atlas sonore des langues régionales de France [14]. All materials were presented by PowerPoint.

The speakers were recorded in Papeete in a quiet room at the Université de la Polynésie Française in November 2018 using a Zoom H6 recorder through a Countryman ISOMAX head-mounted microphone. Recordings had a sampling rate of 44.1kHz and 16 bit quantisation. Speakers produced four repetitions of each token in EXP1 and two or three repetitions of the fable in the BES corpus. Table 2 shows the distribution of glottal and oral stops across the two experimental tasks investigated in this study.

Table 2. *Number and distribution of /ʔ/ and combined oral stops /p/ /t/ according to experimental task and word position*

	EXP1		BES	
	Initial	Medial	Initial	Medial
/ʔ/	377	701	302	306
Oral stops	1502	1057	484	209

2.2. Data processing and annotation procedures

The audio files were forced aligned using a version of the general WebMaus protocol [15] that has been adapted for French Polynesian language inventories, and vowel and consonant boundaries were subsequently checked and hand corrected using the waveform and spectrogram [16]. Additional annotation of both datasets was undertaken to identify the different types of glottal stop variants adapting the fine-grained annotation procedures in [12] for Hawaiian. Glottal stop variants were assigned to different categories ranging from the most consonantal to least consonantal (i.e. more vocalic) as follows:

Full glottal closure (FGC)
Mid creak (creaky phonation during most of the C interval)
Whole creak (creaky phonation across a VCV sequence)
Intensity dip (reduced intensity relative to surrounding Vs)
Breathy phonation
Modal phonation

Where there was a period of voicelessness with some pulses at either the beginning or end of the closure interval, these were identified as Full Glottal Closure (FGC). In many cases a portion of the following vowel was also included if there was a period of creaky voice at the beginning of the vowel. Two examples of /ʔ/ with complete glottal closure (i.e. canonical glottal stops) are shown in Figure 1. Cases of /ʔ/ realised as non-modal phonation (mid creak, whole creak, breathy voice) were also visually identified from waveforms and spectrograms. Cases of creaky phonation showed irregular and damped glottal pulses and breathy voice showed high frequency noise components (after [7]).

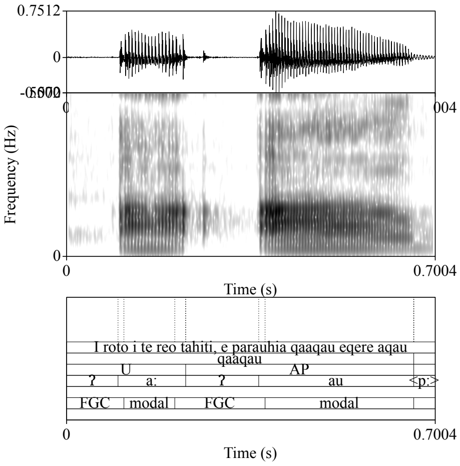


Figure 1: *Waveform and spectrogram of the token /ʔaa'ʔau/ 'conscience/soul' showing two examples of /ʔ/ produced with full glottal closure (FGC) in the EXP1 task.*

Mid creak variants were noted if /ʔ/ was produced with a period of intermittent creaky voice with no real evidence of full glottal closure. Whole creak realisations showed creaky voice across the entire VʔV sequence with no evidence of devoicing. Instances of whole creak were more likely to occur when the flanking vowels were identical. Intensity dip realisations were labelled if /ʔ/ was realised as a period of modal phonation with reduced amplitude. Initial annotations were conducted by the second author and checked by the first author.

2.3. Acoustic analysis procedures

Duration values were extracted for consonant phonemes (separately for glottal and oral stops) and for each annotated glottal variant using the emu-sdms system in R [16]. The oral stop durational values were calculated to enable comparison with glottal stop duration. Similar to [11], we included closure interval plus post-release phases in the overall C measures for voiceless stops. For this reason, only consonants that were not preceded by a pause were included in this analysis. Durational values for each variant type were also calculated after [11,12].

Following [10], VoiceSauce [17] implemented in Matlab was used to extract a number of voice quality parameters including harmonics-to-noise ratio < 1500 Hz (HNR15) which is a measure of inharmonic noise. Creaky phonation is generally associated with lower ratio values compared to modal phonation. Values were extracted for all labelled /ʔ/ and /p, t/ across the two experimental tasks (EXP1 and BES). Other parameters were also extracted including spectral tilt measures (H1*-H2*), and cepstral peak prominence (CPP) but these will be investigated in a forthcoming study.

2.4. Statistical Analysis

The duration data were analysed with linear mixed effects models (LMM) using lmerTest [18] in R [19] with fixed effects consonant category (oral, glottal) and experimental task (EXP1, BES) plus interactions. Preliminary statistical analysis showed that prosodic prominence was not significant so this was not included in the final model. Random effects were included for speaker and word. The voicing measure (HNR15) was submitted to a generalised additive mixed model (GAMM) using mgcv [20] and itsadug [21] in R. The model included parametric difference terms for experimental task and position, smooths over normalised glottal stop duration for position and experimental task, and random smooths by participant and experimental task and word and experimental task. Basis terms were set at k=11 and an AR1 error term was also included in the model to take into account autocorrelation, after [22].

3. Results

3.1. Distribution of glottal variants

Figures 2 and 3 show the distribution of major glottal stop variants plotted separately for experimental task. In this section descriptive rather than inferential statistics are presented to allow comparison with similar studies [8,11,12]. In both experimental tasks, the dominant variant of /ʔ/ has full glottal closure with 77% of all variants in EXP1 (Figure 2) and 59% in the BES corpus (Figure 3) so experimental task is an important factor determining the occurrence of canonical glottal stops. A higher proportion of creaky voiced (i.e. mid creak, whole creak) and other variants (breathy and modal) are realised in the BES task compared to the more controlled EXP1 task. In EXP1, 96% of word-initial variants are canonical glottal stops compared to 65% in word-medial position. It should also be pointed out that there are more than twice as many word-medial glottal variants in EXP1 compared to initial /ʔ/ overall as shown in Table 2. In the BES task, around 66% of word initial glottal stops are produced with full glottal closure with 50% of medial /ʔ/ realised as canonical glottal stops. While creaky voiced variants (i.e. mid creak, whole creak) are produced in medial contexts in both tasks, more whole creak variants (i.e. with creaky voice realised across the entire VCV sequence) and modal voiced realisations are observed in medial position in the BES corpus. Only a handful are produced by the same speakers in EXP1. There are more identical vowel VʔV sequences in the BES corpus which is a prime location for either whole creak or modal realisations according to [4]. Moreover most medial modal realisations in EXP1 are found in the minimal pair /roʔa/ *ro'a* "heartwood" and /roːʔaː/ *rō'ā* "shrub used for fishing lines" which is produced by two speakers without any glottalisation. There are only a handful of intensity dip and modal or breathy variants of /ʔ/ produced by speakers compared to full glottal closure variants.

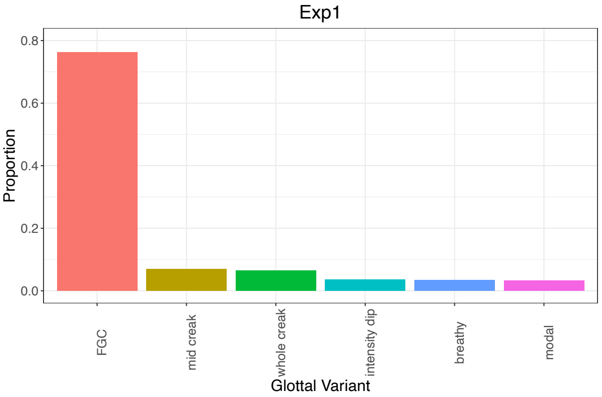


Figure 2: *Distribution of glottal variants of /ʔ/ in Experiment 1*

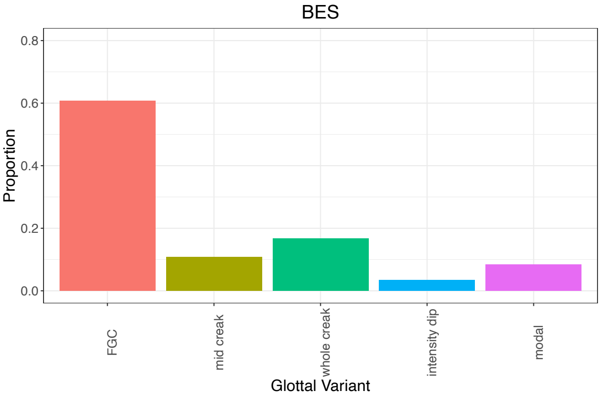


Figure 3: *Distribution of glottal variants of /ʔ/ in the BES corpus*

3.2. Duration

Figure 4 shows a boxplot summarising the simplified distribution of acoustic duration values for all phonemic glottal stops /ʔ/ compared to the combined oral stops /p,t/. There were significant effects of stop type (oral versus glottal) and experimental task on consonant duration, with no interaction between the two factors (consonant type: F=133.28 p<0.0001; task: F= 4.39, p<0.05). Oral stops are significantly longer than glottal stops in general by around 60 ms. Stops are also generally longer in the controlled experimental task compared to the read task.

Figure 5 plots the durational distribution of the main variants of /ʔ/ in both experimental tasks collapsed across word position for ease of visualisation. Recalling the skewed distribution of full glottal closure (FGC) variants across the two tasks from 3.1, only descriptive statistics are presented here. For the most part, variants are longer in EXP1 compared to the BES corpus. The dominant FGC category is the longest variant with a mean duration value of 122 ms in EXP1 and 61 ms in BES. Mid creak variants are also much shorter in the BES corpus compared to EXP1 (58 ms versus 98 ms). In sum, creaky voice variants are shorter than more "consonantal" variants overall. As noted above, the plot also reflects the important differences between the two tasks noted previously with EXP1 showing some variants with modal phonation in particular lexical items, whereas BES contains more variants with whole creak realisations given the higher instance of VʔV sequences where the flanking vowels are identical.

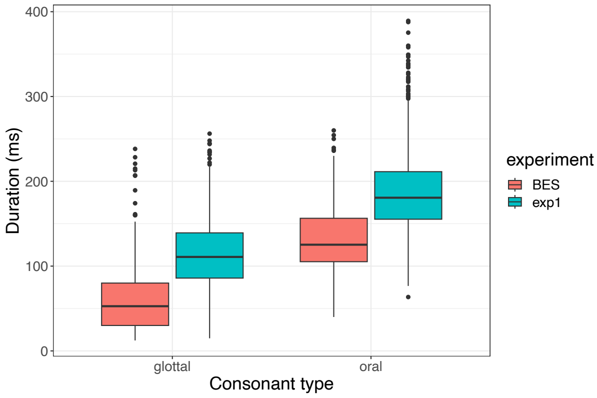


Figure 4: *Glottal stop /ʔ/ and combined oral stop /p,t/ duration plotted by experimental task and word position*

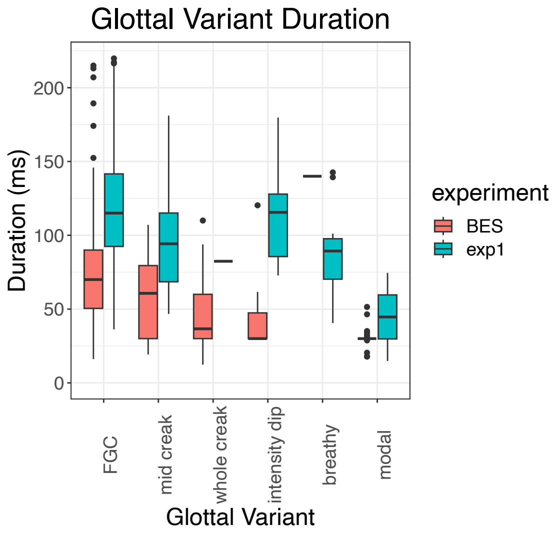


Figure 5: *Acoustic duration of /ʔ/ variants plotted according to experimental task.*

3.3. Voice quality measure: HNR

The time course of harmonics-to-noise ratio is shown in Figure 6 for FGC realisations given this was the main /ʔ/ variant across both tasks. Experimental task was a strong predictor of HNR level with lower HNR values observed in the more constrained experimental task reflecting higher levels of full or partial glottal closure and different trajectory shape (parametric: t=-5.88, p<0.0001, non-linear:F=5.8, p<0.0001). Position was also significant (parametric:t=5.45 (non-linear: F=9.2, p<0.0001) with medial /ʔ/ having higher overall HNR values and a different HNR trajectory shape than initial /ʔ/ as shown in Figure 6.

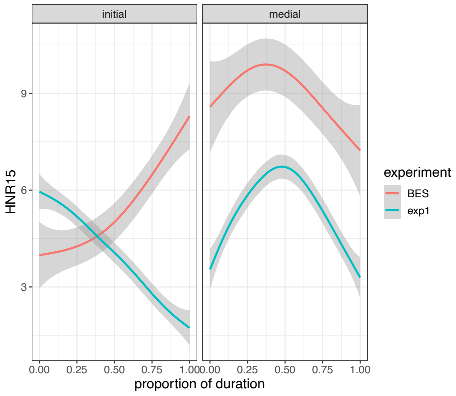


Figure 6: *Loess-smoothed plot of the time course of harmonics-to-noise ratio (HNR) for all /ʔ/ realised with full glottal closure*

4. Discussion

As predicted, /ʔ/ has a number of different glottal variants in Tahitian echoing findings for a range of other languages of the world which also have phonemic glottal stops [8]. Moreover there are different distributions of variants according to experimental task. More canonical glottal stops are produced in the controlled token experiments as a proportion of all glottal variants. Many /ʔ/ variants in the BES corpus, whilst showing a period of voicelessness (i.e. FGC variants) often also show partial closure reinforcing the highly gradient nature of glottal stop production [8,9,11,12]. Compared to Hawaiian, the proportion of canonical glottal stops in both tasks is somewhat higher, with 59% in our BES corpus and 77% in the more controlled experiment compared to 7% reported in [12].

In general, glottal stops in Tahitian tend to be more "consonantal" rather than "vocalic". Even mid-creak realisations tend to be more "consonantal" compared to Hawaiian mid-creak variants [2] particularly in the more controlled task. In terms of acoustic duration, oral stops are longer than glottal stops and both glottal and oral stops are longer in the more controlled experimental task which is not particularly surprising given well documented task and contextual effects on acoustic duration of segments[23]. FGC variants tend to be longer than other variants, particularly in the controlled task, suggesting gestural strengthening, and there are relatively low harmonic-to-noise ratio values similar to the Arapaho canonical /ʔ/. High proportions of true glottal stops are also observed in initial and medial position in the controlled task with the less constrained task showing more non-canonical realisations particularly in medial position, as observed by [8,12]. More vowel-like whole creak and modal variants are observed when /ʔ/ is flanked by identical vowels similar to Hawaiian [12] and as observed by [13].

In sum, there is a degree of expected gradience in the realisation of /ʔ/ in Tahitian together with task-specific inter- and intra-speaker variation (e.g. see 3.1) but at somewhat different levels compared to studies of other languages. Future analyses will examine this further and will include other measures of voice quality including spectral tilt, glottal strength of excitation, and cepstral peak prominence.

Table of Contents for this manuscript

5. Acknowledgements

Aspects of this research were supported by the Australian Research Council Centre of Excellence for the Dynamics of Language CE1401000 and a mission de recherche award from the Université de la Polynésie Française to the first author. Particular thanks are expressed to Jacques Vernaudon and Mirose Paia for assistance with subject recruitment in Papeete. Ben Volchok and Eleanor Lewis also contributed to material preparation and data processing of earlier versions of the corpus.

6. References

[1] Walworth, M. “Eastern Polynesia”, in E. Adamou and Y. Matras. The Routledge Handbook of Language Contact, pp 1-17. 2020.

[2] Paia, M. and J. Vernaudon.. “Le Tahitien : plus de prestige, moins de locuteurs” *Hermès, La Revue*, 2002/1, no 32-33, p. 395-402. 2002.

[3] Bodin, V. “Tahiti:la langue et la société”. Éditions ‘Ura. 2006.

[4] Lemaître, Y. “Lexique du Tahitien contemporain”. Paris:Éditions Orstom 1995.

[5] Perini,A-D. “La phonologie du tahitien et l'intégration des emprunts”. *La Linguistique*, 23 (2), 131-142, 1987.

[6] Blust, R. “*t to k: An Austronesian Sound Change Revisited”. *Oceanic Linguistics* 43 (2), 365-410. 2004.

[7] Ladefoged, P. and Maddieson, I. “Sounds of the world’s languages”. Oxford: Wiley-Blackwell. 1987

[8] Garellek , M, Chai, Y., Huang, Y., and van Doren, M. “Voicing of glottal consonants and non-modal vowels. *Journal of the International Phonetic Association* 53, 305-332. 2021.

[9] Kohler, K. “Glottal stops and glottalization in German”. *Phonetica*, 51, 38-51.1994

[10] Garellek, M. “Voice quality strengthening and glottalization” *Journal of Phonetics 45*(1): 106–113.2014.

[11] Whalen, D., DiCanio, C. and Chen, W-R. “Variable realization of the Arapaho glottal stop despite its being distinctive and frequent”. Proceedings of ICPHS2023, 3276-3280. 2023.

[12] Davidson, L. “Effects of word position and flanking vowel on the implementation of glottal stop: Evidence from Hawaiian”. *Journal of Phonetics* 88, 1-14. 2021.

[13] Lemaître, Yves. “La Phonologie du Tahitien”. Unpublished PhD Manuscript .1971

[14] Boula de Mareüil, P. Vernier F., Adda G., Rilliard, A, Vernaudon, J. “A speaking atlas of indigenous languages of France and its Overseas Territories”, International Conference Language Technologies for All (LT4All), Paris (pp. 155–159). 2019.

[15] Kisler, T., Reichel, U., Schiel, F. “Multilingual processing of speech via web services”. *Computer, Speech, and Language*, 45, 326-347. 2017.

[16] Winkelmann, R., Harrington, J., Jänsch, K. “EMU-SDMS: Advanced speech database management and analysis in R. *Computer, Speech and Language* 45, 392-410. 2017.

[17] Shue, Y.-L., P. Keating , C. Vicenik, K. Yu. “VoiceSauce: A program for voice analysis”, Proceedings of the ICPhS XVII, 1846-1849. 2011.

[18] Kuznetsova A, Brockhoff P.B., Christensen R.H.B.“lmerTest Package: Tests in Linear Mixed Effects Models.” Journal of Statistical Software, 82(13), 1-26. 2017

[19] R Core Team. “R: A language and environment for statistical computing”. R. Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org. 2021.

[20] Wood, S. N. “ Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models.” *J Royal Statistical Society*. 73, 3-36. 2011.

[21] van Rij, J., Wieling, M., Baayen, R.H., van Rinj, H. “Itsadug: interpreting time series and autocorrelated data using GAMMS”. https:/CRAN-R-project.org/package=itsadug. 2022.

[22] Sóskuthy, M. “Evaluating generalised additive mixed modelling strategies for dynamic speech analysis”. *Journal of Phonetics*, 84 https://doi.org/10.1016/j.wocn.2020.101017.2021

[23] Fletcher, J. “The prosody of speech: timing and rhythm”. Handbook of the Phonetic Sciences, 2nd edition, Cambridge, Mass: Wiley-Blackwell, 523-602. 2010

An Alternative Approach to Depression Diagnosis: Predicting Individual Symptoms Through Speech and Text Analysis

Karim M. Ibrahim, Antony Perzo, Larsen D'hiet

Emobot, France

`research@emobot.fr`

Abstract

We propose a symptom-based approach to depression diagnosis by predicting individual symptoms through speech and text analysis. By combining traditional acoustic features with emotion recognition metrics—specifically, valence and arousal from speech and text—our model enhances transparency and accuracy in assessing depression severity [1]. Using the DAIC-WOZ dataset, we demonstrate improved prediction of individual symptoms compared to traditional methods. The clear distinction in valence across depression severity levels underscores the utility of the proposed features. Our method can be seamlessly integrated into existing clinical workflows, offering clinicians a non-invasive and interpretable tool for diagnosis and monitoring. Future work will focus on incorporating additional non-speech-related symptoms and validating the approach in clinical settings to further enhance its applicability and effectiveness.

Index Terms: depression, diagnosis criteria, clinical usefulness, emotion recognition

1. Introduction

Major Depressive Disorder (MDD) is a widespread condition with significant societal and economic impacts, driving the need for innovative approaches to its diagnosis and treatment [2, 3]. Among various digital health initiatives, speech has emerged as a promising indicator due to its hierarchical structure and potential to reflect mental health conditions [4]. Research has shown that speech from individuals with depression exhibits acoustic changes, such as reduced speech rate, decreased pitch variability, and changes in energy, which correspond to key depressive symptoms like diminished emotional expression and psychomotor retardation [5, 6].

Despite these findings, the clinical adoption of speech-based systems for depression diagnosis has been limited. Key challenges include the suboptimal performance of current models, small dataset sizes that limit generalizability, and a lack of transparency, which hinders clinical trust in such tools [1, 7]. Addressing these issues requires an approach that not only improves accuracy but also offers clear insights into the reasoning behind predictions.

In this study, we propose a method that predicts individual symptoms of depression through the analysis of speech and text features. Our approach leverages traditional acoustic features along with symptom-based features, such as valence and arousal, derived from emotion recognition and sentiment analysis [8]. By focusing on individual symptoms rather than a binary diagnosis, we aim to enhance transparency and provide clinicians with detailed, interpretable insights that can improve treatment strategies.

The paper is structured as follows: in section 2, the DAIC-WOZ dataset design is presented, explaining the collection procedure and the depressive symptoms included through the PHQ-8 Questionnaire [9], along with the preprocessing steps for feature extraction. In section 3, we explain our methodology for combining established acoustic measures with proposed symptom-based features, particularly valence and arousal from speech and text, aiming at predicting the depressive symptoms. Finally, in section 4, we present the experiments and evaluation results obtained using our proposed approach.

2. Dataset

Our analysis is built on the DAIC-WOZ dataset, obtained from the USC Institute for Creative Technologies[1]. This dataset, comprising 189 participants, is derived from the larger Distress Analysis Interview Corpus (DAIC), extensively discussed in existing literature [10, 11]. Its primary purpose is to diagnose psychological distress conditions, such as depression, using computer-agent-conducted interviews. The duration of interaction sessions varies from 7 to 33 minutes, averaging at approximately 16 minutes per session. The interviews have been transcribed and annotated within the dataset. Alongside interview transcripts, the dataset provides PHQ-8 scores, a widely employed diagnostic measure for depression assessment.

The dataset includes the answers for each of the individual questions in the PHQ-8 questionnaire. Each of these questions relate to one of the MDD symptoms, namely: 1) lack of Interest, 2) sadness, 3) sleep disruption, 4) tiredness, 5) appetite disruption, 6) feelings of failure, 7) troubles concentrating, and 8) psychomotor retardation. Each question has one of four answers about the frequency of the symptom: 1) Not at all, 2) several days, 3) more than half the days, and 4) nearly everyday. Our objective is to predict these symptoms independently using tailored features.

The PHQ-8 score distribution within the dataset, illustrated in Figure 1, reveals a predominance of sub-clinical and mild depressive symptoms, with a ratio of approximately 3:1 favoring non-depressed over depressed individuals. This distribution pattern suggests an inherent challenge in achieving high sensitivity in model predictions, given the lower prevalence of depressive symptoms among the participants. Additionally, the PHQ-8 scores are split into four categories to represent the severity: 1) below 5 is classified as `low depression severity`, 2) 5 to 10 as `mild depression`, 3) 10 to 15 as `moderate depression`, and 4) above 15 as `severe depression`. These are the four categories we use for classi-

[1] USC Institute for Creative Technologies, http://dcapswoz.ict.usc.edu/

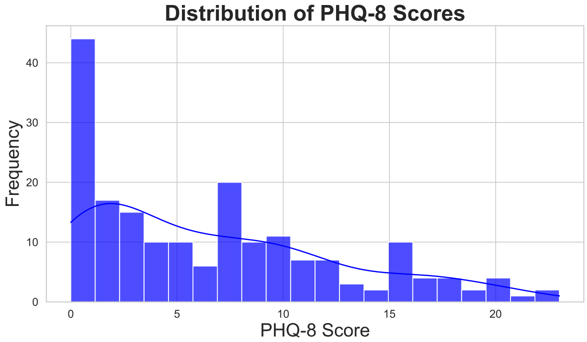


Figure 1: *The PHQ-8 score distribution within the DAIC-WOZ dataset. The distribution is highly skewed towards subclinical cases.*

fication throughout this study.

One of the main challenges in this dataset is the presence of the interviewer's speech in some of the recordings. We address this by applying a pre-processing step of speaker diarization to extract the patient's excerpts. We rely on using the open-source toolkit Pyannote [12, 13] to achieve this. We first apply speech segmentation on the entire interview to extract the speech excerpts. Then, we apply speaker identification to separate between the interviewer and interviewee. We select the most common speaker in the interview to be assigned as the patient, and extract the acoustic and textual features for the corresponding excerpts.

The dataset is divided into training (107 individuals) and development (35 individuals) sets to aid in the evaluation of machine learning models. However, to test the generalizability of our findings across the dataset, we implemented 5-fold cross-validation on the entire dataset such that each validation fold contained at least one sample for every PHQ-8 score in order to counter inherent sampling bias present in the dataset. We used the created splits to train and assess various classifiers, namely Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF). Out of which, SVM provided the best results, which are presented thereafter.

3. Methodology

We utilize both traditional acoustic features and novel symptom-based features to predict depressive symptoms from speech and text. This section outlines the key components of our feature extraction process and the machine learning models used for classification.

3.1. Acoustic features

We rely on the eGeMAPS feature set [14], a standardized set of acoustic features developed to capture emotional expression and affective states in speech. It includes descriptors related to frequency, energy, and spectral balance. To complement this, we include the first 13 Mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives, which are widely used in speech emotion and depression detection [15, 16]. All features are aggregated using the mean and standard deviation across the patient's speech excerpts from the interviews.

3.2. Symptoms-based features

In addition to traditional acoustic features, we propose several symptom-based features designed to capture long-term markers of depression, such as diminished emotional expression and psychomotor retardation [8]. These features focus on sentiment analysis in speech and text, as well as speech rate.

Speech-based sentiment: One core symptom of depression is a persistent low mood and loss of interest or pleasure (anhedonia) [8]. Recent advancements in speech analysis have provided opportunities to detect depression through short-term emotional states, such as valence (degree of pleasantness) and arousal (level of excitement) [5, 17]. Individuals with depression exhibit distinct emotional reactivity patterns, which can be effectively captured through speech, making emotional state tracking a valuable tool for early detection and personalized treatment [18, 19]. Using Speech Emotion Recognition (SER) systems, specifically pretrained models like Wav2vec 2.0, has shown significant promise for estimating valence and arousal from speech [20, 21]. Wav2vec 2.0 employs a convolutional neural network (CNN) to encode raw audio into low-level representations, followed by transformer layers to capture contextual information, and an optional quantization module to discretize the representations into units as needed for certain tasks [20]. In this study, we fine-tune Wav2vec 2.0 for the SER tasks, as suggested by approaches such as [22, 23], and use the MSP-Podcast dataset for fine-tuning [24]. The model achieves strong performance with concordance correlation coefficients (CCC) of 0.635 for valence and 0.745 for arousal on the MSP-Podcast test set [25]. After processing patient speech excerpts, we compute the mean and standard deviation for valence and arousal across all excerpts to correlate these features with depressive symptoms.

Text-based sentiment: Another important indicator of depression is the content and choice of words in speech [8]. To compute the valence and arousal values from patient transcripts in the DAIC-WOZ dataset, we use the NRC Valence, Arousal, and Dominance (VAD) Lexicon [26], a comprehensive resource containing emotional ratings for around 20,000 English words. The NRC VAD Lexicon was developed to facilitate emotion analysis, sentiment analysis, and related fields, helping to discern emotional content from textual data. Its ratings are based on human judgments collected via crowdsourcing, ensuring that the emotional associations are grounded in widespread human perception. This makes it a valuable tool for research across various domains, including customer feedback analysis, social media monitoring, and psychological studies. For each transcript excerpt, we extract the valence and arousal values for individual words and calculate the average values across the excerpt. We then compute the mean and standard deviation of these values across the entire interview to capture the patient's overall emotional expression.

Rate of speech: One of the main symptoms of depression is moving and speaking more slowly [8]. We aim at measuring this symptom using 2 features: *The total number of words in an utterance* and *the rate of speech*, which is calculated as the mean of the number of words divided by the length (in seconds) of each utterance. The unit is thus words/s. Utterances shorter than 7 words, chosen empirically, are not taken into account for a more reliable estimate. Similar to the previous features, we compute the mean and standard deviation across the whole interview for the patients' excerpts.

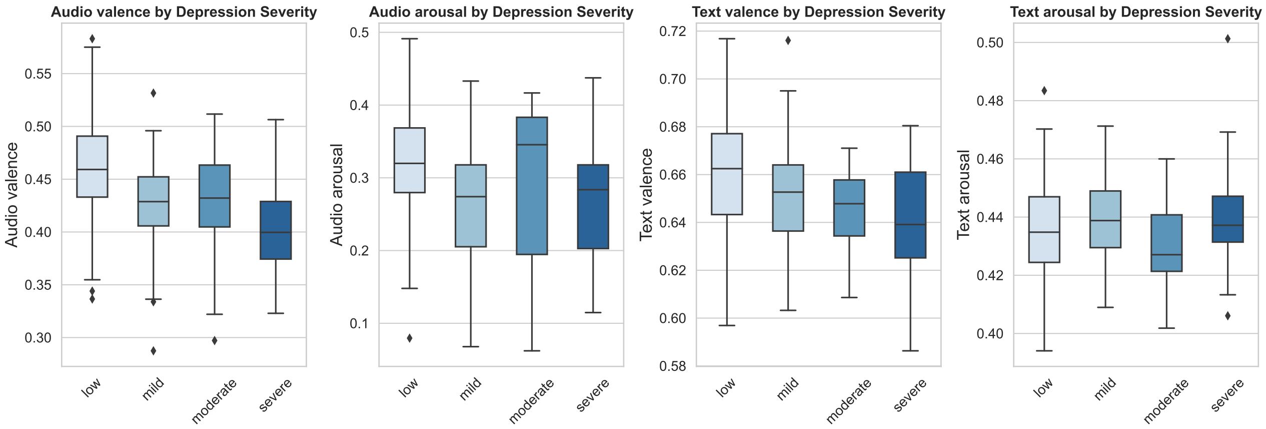


Figure 2: *Distribution of the speech and text valence values across different depression categories. We observe a clear distinction between low depression severity and severe depression in the values of the valence.*

3.3. Model setup

To evaluate the effectiveness of these features, we use a Support Vector Machine (SVM) with a radial basis function (RBF) kernel as our classification model. The eGeMAPS features, combined with the proposed symptoms-based features, are used as input to the model. Features are standardized prior to training, and feature selection is applied to reduce dimensionality and enhance performance. The number of selected features and the SVM's regularization factor are tuned as hyperparameters to optimize the model's accuracy and generalization. This approach enables us to assess the contribution of both traditional acoustic and symptoms-based features in predicting depressive symptoms.

4. Results and Discussion

4.1. Distribution analysis

First, we examine the distribution of valence and arousal in both speech and text across varying levels of depression severity, as illustrated in Figure 2. The valence values (on a scale from -1 to 1, where lower values indicate more negative emotions) show a consistent decrease as depression severity increases. This trend is particularly noticeable in the distinction between low depression severity and severe depression, where individuals with higher depression scores exhibit notably lower valence values in their speech. This finding suggests that as depression becomes more severe, the emotional content of speech shifts toward more negative expressions. Similarly, the text valence values also drop with increasing depression severity, indicating that participants use more negatively connoted language as their depression worsens. This supports the idea that speech and text analysis can effectively capture emotional states linked to depression. In contrast, the arousal values (measured on a scale from 0 to 1, where lower values indicate calmer or less excited speech) present a more overlapping distribution across different depression severity levels. While there is some variation in arousal, the overlap suggests that arousal may be less strongly correlated with depression severity compared to valence. This could be because arousal reflects a broader range of emotional and physiological states that may not be as tightly linked to depression as valence.

4.2. Performance evaluation & Feature importance analysis

We evaluated the performance of the models by comparing two sets of features: acoustic features alone and a combination of acoustic and symptoms-based features. Table 1 presents the results for predicting the severity of each depression symptom, with metrics reported as mean ± standard deviation across a 5-fold cross-validation process. The metrics used are accuracy, recall, precision, and F1-score, with all values scaled between 0 and 1, where higher scores indicate better model performance.

The F1-score is a performance metric that combines precision (the proportion of true positive predictions among all positive predictions) and recall (the proportion of true positive predictions among all actual positive cases) into a single score. It is particularly useful when the data is imbalanced, as it provides a balance between precision and recall, helping to assess the model's effectiveness in cases where false positives and false negatives are both important.

The inclusion of symptoms-based features led to improvements in nearly all metrics across various symptoms. Overall, accuracy in detecting specific symptoms using only acoustic features ranged from 0.34 in sleep disruption to 0.71 in psychomotor retardation, whereas the combined feature set allowed the accuracy for psychomotor retardation to rise to 0.75. Similarly, the F1-score for depression severity classification improved from 0.32 to 0.36. These results demonstrate that integrating emotional and symptom-specific features with traditional acoustic features enhances the model's ability to capture the nuanced patterns associated with different depression symptoms.

In contrast, for symptoms such as sleep disruption and feelings of failure, the improvements were more modest, and in some cases, performance slightly decreased. For example, accuracy for sleep disruption remained at 0.34, and the F1-score showed no significant improvement in these cases. This may indicate that speech-based features alone may not fully capture certain symptoms, particularly those that are not directly observable through vocal characteristics, such as physical symptoms like sleep and appetite disruptions.

The results also highlight the importance of valence and arousal in symptom prediction. The clear distinction in valence

Table 1. *Evaluation results on the test set for predicting the severity of each symptoms of depression using acoustic features and proposed features. Evaluation using accuracy , recall, precision, and F1-score. results are presented as mean* ± *std. across 5-fold cross validation.*

	Acoustic Features				Acoustic + Symptoms-based Features			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
Lack of Interest	0.44 ± 0.11	0.28 ± 0.08	0.27 ± 0.07	0.27 ± 0.07	0.42 ± 0.13	0.30 ± 0.09	0.31 ± 0.10	**0.30 ± 0.10**
Sadness	0.41 ± 0.06	0.34 ± 0.12	0.37 ± 0.15	0.34 ± 0.13	0.42 ± 0.05	0.36 ± 0.08	0.40 ± 0.05	**0.35 ± 0.04**
Sleep Disruption	0.34 ± 0.06	0.33 ± 0.08	0.35 ± 0.05	**0.33 ± 0.06**	0.35 ± 0.09	0.32 ± 0.05	0.33 ± 0.04	0.31 ± 0.05
Tiredness	0.35 ± 0.06	0.27 ± 0.06	0.29 ± 0.04	0.26 ± 0.04	0.37 ± 0.04	0.30 ± 0.05	0.38 ± 0.11	**0.30 ± 0.06**
Appetite Disruption	0.38 ± 0.08	0.28 ± 0.03	0.27 ± 0.04	0.26 ± 0.03	0.35 ± 0.06	0.30 ± 0.04	0.29 ± 0.03	**0.29 ± 0.03**
Feelings of Failure	0.40 ± 0.08	0.31 ± 0.10	0.38 ± 0.14	**0.31 ± 0.11**	0.39 ± 0.08	0.30 ± 0.02	0.32 ± 0.04	0.30 ± 0.03
Concentrating Troubles	0.39 ± 0.02	0.33 ± 0.11	0.30 ± 0.04	0.30 ± 0.05	0.51 ± 0.05	0.32 ± 0.07	0.32 ± 0.10	**0.31 ± 0.08**
Psychomotor Retardation	0.71 ± 0.06	0.31 ± 0.07	0.30 ± 0.10	0.30 ± 0.07	0.75 ± 0.04	0.34 ± 0.04	0.39 ± 0.08	**0.34 ± 0.04**
Depression Severity	0.41 ± 0.09	0.35 ± 0.10	0.36 ± 0.06	0.32 ± 0.09	0.49 ± 0.07	0.38 ± 0.07	0.39 ± 0.12	**0.36 ± 0.07**

across depression severity levels (especially between low and severe categories) suggests that valence is a strong indicator of depression severity. On the other hand, the limited differentiation in arousal implies that it may be less effective as a standalone predictor, though it likely contributes to the overall emotional profile when combined with other features.

4.3. Limitations and future work

While the results are promising, there are important limitations to consider. One of the key challenges lies in the reliance on self-reported PHQ-8 scores, which, while widely used in clinical settings for depression screening, may introduce biases. Individuals may underreport or overreport their symptoms due to recall bias or difficulties in recognizing and articulating their emotional states. This self-assessment process may not fully capture the complexity of depression, particularly in cases where symptoms fluctuate over time or where patients experience difficulty in emotional awareness. The PHQ-8's focus on the frequency of symptoms over the past two weeks may also fail to account for individuals with high variability in their symptoms, potentially overlooking critical nuances in their mental health. Moreover, the PHQ-8 questionnaire predominantly addresses cognitive and affective symptoms, such as mood, interest, and concentration, while placing less emphasis on physical symptoms like sleep disturbances and appetite changes, which are crucial aspects of depression for some individuals. As a result, it may not provide a complete picture of the disorder for those whose depression manifests primarily through physical symptoms. [8]

Additionally, the dataset used in this study primarily emphasizes speech-related symptoms, which may leave out other crucial non-speech-related symptoms, such as sleep disturbances and appetite changes, that play an integral role in understanding depression comprehensively. These symptoms are not directly observable through speech analysis, yet they are essential indicators of mental health and significantly influence diagnosis and treatment planning.

Future research should aim to address these limitations by adopting multimodal approaches that incorporate not only speech features but also physiological markers (e.g., heart rate variability, sleep patterns) and behavioral data (e.g., movement tracking, social interaction metrics). Such multimodal data would improve the robustness and accuracy of depression diagnosis models by capturing a more holistic view of the patient's mental state. Expanding the scope of analysis to include these non-speech symptoms would enrich our understanding of depression and enhance the precision of diagnostic tools, ultimately leading to more tailored and effective treatment interventions.

5. Conclusion

The research presented in this paper demonstrates the potential of a symptom-based approach to depression diagnosis, which focuses on predicting individual depressive symptoms rather than providing an overall binary classification. By leveraging a combination of traditional acoustic features and emotionally-driven features such as valence and arousal, we achieved more accurate and interpretable predictions of specific symptoms of depression. In particular, the clear differentiation in speech valence between low and severe depression severity highlights the utility of these features in identifying emotional states associated with depressive disorders.

This approach addresses some of the challenges associated with current depression diagnosis methods, including the need for greater transparency and clinical interpretability. By offering insight into specific symptoms, clinicians can better tailor their interventions, leading to more personalized and effective treatment strategies.

However, this study also underscores the need for further exploration. The integration of non-speech-related symptoms, such as sleep patterns and appetite changes, will be essential for developing a more comprehensive diagnostic tool. Future research should focus on multimodal approaches that incorporate not only speech and text analysis but also physiological and behavioral data to capture a broader range of depression symptoms. Such developments could significantly enhance the precision and clinical usefulness of depression diagnosis systems.

In conclusion, this study offers a promising direction for improving the assessment of depression through the prediction of individual symptoms, contributing to the broader goal of developing transparent, interpretable, and clinically applicable diagnostic tools in mental health.

6. References

[1] V. P. Martin and J.-L. Rouas, "Estimating symptoms and clinical signs instead of disorders: the path toward the clinical use of voice and speech biomarkers in psychiatry," in *Proceedings of the International Conference on Acous-*

tics, Speech, and Signal Processing (ICASSP), 2024.

[2] F. Matcham, C. Barattieri di San Pietro, V. Bulgari, G. De Girolamo, R. Dobson, H. Eriksson, A. Folarin, J. Haro, M. Kerz, F. Lamers *et al.*, "Remote assessment of disease and relapse in major depressive disorder (radar-mdd): a multi-centre prospective cohort study protocol," *BMC psychiatry*, vol. 19, pp. 1–11, 2019.

[3] C. G. Fairburn and V. Patel, "The impact of digital technology on psychological treatments and their dissemination," *Behaviour research and therapy*, vol. 88, pp. 19–25, 2017.

[4] N. Topooco, H. Riper, R. Araya, M. Berking, M. Brunn, K. Chevreul, R. Cieslak, D. D. Ebert, E. Etchmendy, R. Herrero *et al.*, "Attitudes towards digital treatment for depression: a european stakeholder survey," *Internet interventions*, vol. 8, pp. 1–9, 2017.

[5] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[6] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[7] M. Milling, F. B. Pokorny, K. D. Bartl-Pokorny, and B. W. Schuller, "Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell," *Frontiers in digital health*, vol. 4, 2022.

[8] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Publishing, 2013. [Online]. Available: https://books.google.fr/books?id=-JivBAAAQBAJ

[9] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[10] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014.

[11] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014.

[12] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, 2023.

[13] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, 2023.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010.

[16] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *Proceedings of the international conference on acoustics, speech and signal processing (ICASSP)*, 2014.

[17] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[18] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[19] S. Koops, S. G. Brederoo, J. N. de Boer, F. G. Nadema, A. E. Voppel, and I. E. Sommer, "Speech as a biomarker for depression," *CNS & Neurological Disorders Drug Targets*, vol. 22, no. 2, pp. 152–160, March 2023. [Online]. Available: https://doi.org/10.2174/1871527320666211213125847

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, 2020.

[21] P. Dagum, "Digital biomarkers of cognitive function," *npj Digital Medicine*, vol. 1, no. 1, p. 10, March 2018. [Online]. Available: https://doi.org/10.1038/s41746-018-0018-4

[22] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, 2021.

[23] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[24] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[25] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[26] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the 56th annual meeting of the association for computational linguistics*, 2018.

Investigating Word Retrieval Bias During Verbal Semantic Fluency Tasks in Adolescents At-Risk for Mental Health Disorders

Brian Stasak[1,2] *and Mark Larsen*[2,3]

[1]School of Elect. Eng. & Telecomm., University of New South Wales, Australia, [2]Black Dog Institute, University of New South Wales, Australia, [3]Centre for Big Data Research in Health, University of New South Wales, Australia

b.stasak@unsw.edu.au, mark.larsen@unsw.edu.au

Abstract

This study investigates mental health (e.g., anxiety, depression, distress, suicidal ideation) and word-retrieval biases exhibited by 122 adolescents in 188 verbal categorical semantic fluency audio recordings. Traditional metrics show that adolescents with suicidal ideation produced up to twice as many unrelated keywords than those without. Experimental keyword metrics (e.g., affective ratings, dietary nutrients, RGB colour, sport attributes) were also evaluated. Results show that *high* severity groups produce food category keywords that are low in fibre and vitamin C; and *high* severity groups for anxiety and suicidal ideation generate food category keywords with significantly higher caloric, fat, cholesterol, and protein content.

Index Terms: anxiety, depression, distress, speech, suicidality

1. Introduction

Studies [1, 2] have shown that mental illness is pervasive in school aged populations. This is concerning because early episodes of mental illness often result in life-long wellbeing struggles [3]. According to global studies [4, 5], half of mental health illnesses found in adulthood begin by age 14 and most of these cases go undetected and/or untreated. While early screening in teenagers with mental illness improves clinical outcomes, there is still a desire to better understand the complex heterogenous symptoms associated with mental illness and its effects on verbal behaviour in this vulnerable population.

While hundreds of subjective self-survey mental health severity screening protocols exists for mental illnesses like depression (e.g., BDI-II, PHQ-9, HAMD) [6], experimental studies [7–9] have explored objective speech-based methods to further help identify mental illness symptoms. For instance, during isolated verbal tasks, speech-based mental health studies (e.g., bipolar, suicidal behaviour) [7–10] have reported abnormal neurocognitive and physiological speech-language behaviours (e.g., disfluencies, flat prosody, poor voice quality).

While abnormal emotional bias in adults with clinical depression has been previously studied, highlighting a propensity for increased negative valence fixation [11], risk word usage [12], and personal pronouns [13], there has been little investigation into other indices that could reveal verbal word retrieval bias in individuals with mood disorders. One isolated method that could be useful in quickly ascertaining bias is the semantic fluency task, which is a verbal test that gauges how well a person can quickly name items in specific category. Each utterance provides a degree of biased idiolect and can help to reveal mood-related disturbances [12, 13].

Unlike previous verbal semantic fluency mental illness studies [9, 14] which have focused on calculating traditional semantic fluency metrics, such as the number correct, the study herein also investigates category-specific keyword indices, such as affective human ratings, food nutrients, RGB colours, and sport traits. When compared to healthy controls, conversational speech transcripts studies [14, 15] have found lower valence ranges in adults with mood disorders. Therefore, it may be hypothesized that during the semantic fluency emotion category task, adolescents with higher mental health disorder severity scores will utter keywords with lower valence values (i.e., more negative than positive terms).

Previous diet-related mental health studies [16-20] indicate that adolescents diagnosed with mental health disorders often have unhealthy, inadequate diets when compared to healthy controls. For instance, adolescents with clinical depression or suicidal behaviour were shown to consume less fibre [18], less vitamin C [19], and more fat [20]. While deeper insights into nutrimental contents of food keywords uttered by participants have not yet been investigated using semantic fluency tests, it is likely that participants will verbally project their personal eating habits, reflecting the previously mentioned associations.

Studies [21, 22] that have investigated the relationship between colour and specific emotions/moods, have found that colour brightness is associated with positive-emotional words and colour darkness with negative-emotional words. For example, keywords associated with mood disturbances, such as '*anger*', '*failure*', and '*sadness*', are frequently associated with certain culturally designated darker colours, such as red and blue [22]. It is anticipated that during colour semantic fluency tasks individuals with higher mental health disorder severities may more frequently utter darker coloured keywords.

For sport semantic fluency tasks, considerations into competition attributes could also provide signs to mood disorders. It is hypothesized that adolescents with higher symptom severities may prefer to name sports that are individually played or have a smaller number of players; and that avoid direct physical contact with others. A recent study [23] found that sports team athletes were less likely to have anxiety or depression than solo athletes.

This speech-based mental health study examines 122 adolescents' verbal semantic fluency task audio recordings in natural real-world environments to explore differences in responses from participants in *low* (none-to-minimal) versus *high* (moderate-to-severe) severity groups. Using a statistical analysis per severity group, four semantic fluency categories and new index metrics are examined. Based on recorded speech-transcripts, results test the validity of using deeper lexical indices to help identify youths who are more at risk for anxiety, distress, depression, and suicidal ideation disorders.

2. Data

Experiments in this study used speech recordings collected for the Future Proofing Study in Australia [24]. This data included mental health self-report surveys, demographic metadata, and elicited audio recordings collected in Australia using an interactive smartphone app [24]. This data collection study was approved by the University of New South Wales Human Research Ethics Committee. Participants were Year 8 high school students (i.e., 12 to 14 years old) and voluntary enrolment included individual and parental consents. Four semantic fluency task categories, shown in Table 1, were randomly allocated to participants. Recordings were manually vetted for task compliance because some participants failed to comply or there was the presence of excessive background noise. A subset of the original data, herein called BDI-SF, was used for experimentation. In total, there were 122 unique participants and 188 digital audio recordings.

BDI-SF audio was recorded using a mono-channel 44.1 kHz 16-bit sampling rate. Prior to each recording, participants were given read task instructions via a smartphone app. Each recording contained a semantic fluency task (see Table 1), whereby a participant was asked to verbalize in English as many keywords related to a category in a 30-second period. Some participants chose to stop their recordings earlier than the 30-second period. The average recording length was 22 seconds.

Based on mental health self-report surveys (Spence Children's Anxiety Scale Short-Form, Depression Patient Health Questionnaire for Adolescents, Distress Questionnaire-5, Suicidal Ideation Attributes Scale), severity scores were obtained for most participants. Thresholds were used to determine *low* (none-to-minimal) versus *high* (moderate-to-high) severity groups: anxiety (≥10); depression (≥10); distress (≥10); and suicidal ideation (≥1). These cut offs were used in previously published mental health studies [10, 24, 25] and increased sensitivity and sample size per severity group.

Table 1. *Number of BDI-SF dataset participants per low (none-to-minimum) and high (moderate-to-severe) mental health self-survey severity group. More than half of the participants were in the high severity groups for two or more self-surveys.*

	Anxiety		**Depression**		**Distress**		**Suicidal Ideation**	
Task Category	*Low*	*High*	*Low*	*High*	*Low*	*High*	*Low*	*High*
Colours	30	22	31	21	16	36	35	10
Emotions	28	17	27	18	6	39	27	17
Foods	26	16	29	13	17	25	33	9
Sports	30	19	31	18	16	33	33	12
Average Survey Score	*6.6*	*14.2*	*5.2*	*15.7*	*6.8*	*14.4*	*0.0*	*14.5*
Survey Range	*0 to 24*		*0 to 27*		*0 to 25*		*0 to 50*	

3. Experimental Methods

BDI-SF audio recordings were manually transcribed for validation purposes due to variable background noise factors (i.e., automatic speech recognition could be used, but with likely less word recognition accuracy than a human listener). Per recorded transcript, traditional semantic fluency metrics included: the total recording duration in seconds (TRDS); the unique total correct keywords (TCK) related to a given category; percentage of unrelated words (PUW) which was calculated by subtracting TCK from the total number of words then dividing by total number of words; and percentage of word repeats (PWR) which was calculated by counting all word repeats then dividing it by the total number of words uttered.

For the colours category task, the RGB colour code index [26] was used which contained over 200 colour keywords including colour modifiers (i.e., '*burnt sienna*', '*dark blue*', '*neon yellow*'). The RGB has a standardized scale from 0 to 255 for each colour component: (R)ed, (G)reen, and (B)lue. For each of the three colour components, lighter colours are represented by a higher index value, whereas darker colours are represented by a lower index value. Per colour semantic fluency transcript, the RGB component values for each colour keyword uttered were computed. The semantic fluency colour category task transcripts contained 90 unique colour keywords. Fig. 1 contains BDI-SF dataset RGB (R)ed values and corresponding paired colour keyword transcript responses from a *low* and a *high* depression severity participant. These two participants uttered the same number and many of the same colour keywords. However, the *high* severity participant produced a word repeat for the colour '*green*' and generated an average RGB index (R)ed value of 168, whereas the *low* severity participant had a RGB index (R)ed value of 133.

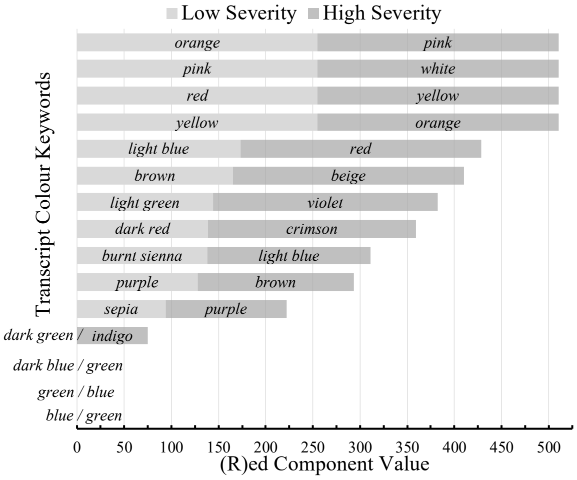


Figure 1: *BDI-SF dataset colour task comparing keyword transcript paired responses from two participants with different depression self-survey severities; (R)ed scale 0-255 per colour.*

For the emotions category task, recording transcripts were evaluated using affective valence keyword human-ratings [27]. The affective valence lookup index included over 20k English words, and each word rating was normalized from 0 to 1 (i.e., a score closer to 1.00 has positive valence, whereas a score closer to 0.00 has more negative valence). In total, semantic fluency emotion category task transcripts contained 147 unique valence-related keywords (e.g., happy, sad, angry, bored).

For the foods category task, a food attribute index was compiled based on an online open-source dietary food guide [28], which contains over 100k food products. Food keywords were based on a serving size of 100 grams and thirteen nutrients were measured: total calories, fat, saturated fat, protein, carbohydrates, sugars, fibre, cholesterol, sodium, iron, potassium, vitamins A and C. In total, participants uttered 208 unique food keywords, including connective terms (e.g., *'bacon and egg sandwich'*, *'spaghetti Bolognese'*).

For the sports category task, four proposed metrics were defined in this study. The sport competition metric included player type: solo-based (-1), mixed (0), and team-based (1). The opponent interaction metric was based on physical opponent contact type: indirect (-1), mixed (0); and direct (1). The location type metric was determined by the location where the sport is typically played: indoor (-1), mixed (0), and outdoor (1). For example, rugby is a team-based (1) direct competition (1) in an outdoor (1) setting, whereas fencing is a solo-based (-1) direct competition (1) in an indoor setting (-1). Per sport, the number of on-field teammates was also calculated (i.e., ranging from 1 to 22). In total, participants generated 110 unique sporting activities during the semantic fluency sports task.

For each index measure described above, an automated python script was created to generate participants' task keyword transcript averages. For experiments herein, there was no instance where a semantic fluency transcript contained a correct keyword that was missing from a topic specific index. A composite score involving multiple semantic fluency task indices was inexecutable because participants did not complete all the same tasks (i.e., randomized). Per participant index results, statistical significance between *low* and *high* severity group results were evaluated using two-sample heteroscedastic unequal variance *t*-tests with two reported confidence levels ($p \leq 0.05$, $p \leq 0.10$) [29] and Hedges' effect size ($g \geq 0.40$) [30, 31]. The *t*-test tests whether the difference between response values for two groups is statistically significant or not. The effect size is a quantitative reflection of the magnitude of phenomena concerning how strong a difference there is between two groups. Hedges' effect size provides a better estimate for smaller sample sizes when compared to other methods (e.g., Cohen's *d*, Glass' Δ) [31].

4. Results

Per *low* and *high* severity groups, traditional semantic fluency differences (i.e., low severity minus high severity averages) are shown in Table 2, wherein a positive value indicates that the *low* severity group had a larger average value than the *high* severity group (i.e., a negative value indicates the *high* severity group had a larger value than the *low* severity group). While the total recording duration (TRDS) between the *low* and *high* severity groups were significantly different for many tasks, it should be noted that this measurement does not account for the total duration of speech versus non-speech. For distress and depression, the *high* severity group demonstrated shorter average recording lengths (18.7sec) per emotion category tasks than the *low* severity group (22.8sec). However, the suicidal ideation self-survey emotion category results were an exception because the *high* severity group exhibited longer average recording lengths (23.2sec) when compared to the *low* severity group (20.1sec). Despite the lengthier recordings in the emotions category *high* suicidal ideation severity group, this extra time did not result in a higher percent average of correct words produced (i.e., this duration increase was due to increased unrelated words). The emotions category TRDS results also show that the *low* severity groups for distress, depression, and suicidal ideation had significantly longer recordings (~3 to ~6 seconds) and were less likely to terminate their recordings early when compared to the *high* severity groups. The *high* severity groups possibly chose to end the emotion category recordings early due to emotional avoidance and/or triggers.

The percentage of unrelated words (PUW) was a strong indicator for the *high* severity depression and suicidal ideation groups during the foods task, whereby the *high* severity group results produced significantly greater proportions of unrelated words (40%, 31%) when compared to the low severity groups (20%, 17%). Generally, the PUW was larger for the *high* severity groups. For different disorders, the *low* and *high* severity groups produced overall higher total correct keywords (TCK) averages for the colors (12.8 to 14.6), foods (12.8 to 13.8) and sports (11.9 to 13.0) tasks, whereas the TCK was lower for the emotions (9.04 to 10.24) task. Thus, hinting that some categories are easier and/or more readily familiar. During the foods task, there was a consistent significant increase in the percentage of word repeats (PWR) for the *high* severity groups for anxiety, distress, depression, and suicidal ideation when compared to the *low* severity groups.

Table 2. *Average difference between low and high severity groups for each semantic fluency category using traditional measurements (recording duration in seconds (TRDS), percentage of unrelated words (PUW), total correct keywords (TCK), percentage of word repeats (PWR)). Shaded results indicate significant group differences based on t-tests ($p \leq 0.05$, $p \leq 0.10$) and medium-to-large Hedges' effect size.*

Self-Survey	Category	TRDS	PUW	TCK	PWR
Anxiety	Colours	-0.36	-6%	-1.40	-2%
	Emotions	-0.17	-9%	-1.04	-7%
	Foods	-1.54	-5%	0.32	-5%
	Sports	-1.12	0%	2.30	0%
Distress	Colours	0.43	-1%	-0.45	-2%
	Emotions	3.73	0%	1.55	1%
	Foods	-0.63	-1%	0.15	-4%
	Sports	0.13	-1%	-1.56	1%
Depression	Colours	-0.88	1%	-2.22	1%
	Emotions	3.09	-16%	0.44	-3%
	Foods	-1.50	-10%	1.44	-7%
	Sports	-0.52	-3%	-0.38	4%
Suicidal Ideation	Colours	-0.56	-5%	-5.10	0%
	Emotions	5.94	-6%	1.06	-1%
	Foods	-1.20	-16%	0.33	-9%
	Sports	1.52	-1%	0.19	5%

Per mood disorder self-survey, the colours semantic fluency results indicated that most average RGB individual component index values were similar for *low* and *high* severity groups. For example, individual component ranges recorded included: (R)ed (142 to 157), (G)reen (123 to 133), and (B)lue (105 to 111). There were two exceptions whereby significant (R)ed average differences for the *low* and *high* anxiety/depression severity groups were recorded. Surprisingly, for the anxiety/depression conditions, results indicated that the *high* severity groups produced (R)ed colour keywords that were brighter in shade (157, 157) than the *low* severity groups (146, 123). Further analysis revealed that the increased (R)ed value found in the *high* severity anxiety/depression groups was attributed to a greater number of red-hued colors uttered per recording (i.e., an average of one more red keyword than *low* severity group). Previous research [21] indicated that the colour red is most associated with negative emotion, which may explain this phenomenon exhibited by the *high* severity groups.

Table 3. *Food semantic fluency category results per low (none-to-minimum) and high (moderate-to-severe) groups. Per food keyword, dietary contents were measured using grams, except for sodium, calcium, and potassium which used milligrams. Shaded results indicate significant group differences based on t-tests ($p \leq 0.05$, $p \leq 0.10$) and effect size.*

		Average													
Self-Survey	**Severity**	**Cals.**	**Fat**	**S. Fat**	**Prot.**	**Carb.**	**Sugar**	**Fibre**	**Chol.**	**Sodi.**	**Calc.**	**Iron**	**Pota.**	**Vit. A**	**Vit. C**
Anxiety	***Low***	186.1	8.2	2.9	7.0	21.9	7.0	2.1	28.6	224.7	47.2	1.4	252.4	47.5	8.5
	High	233.8	11.8	4.8	9.6	23.7	8.4	1.4	39.2	279.3	47.1	1.1	256.8	47.7	7.8
Distress	***Low***	199.0	9.3	3.5	7.7	22.0	7.7	2.4	32.0	248.4	52.3	1.4	251.3	45.6	9.2
	High	192.2	8.6	3.2	7.3	22.4	7.0	1.6	29.7	226.2	43.8	1.2	256.0	49.0	7.6
Depression	***Low***	194.9	8.8	3.2	7.6	22.0	6.5	2.1	29.2	254.1	49.1	1.3	254.6	42.9	8.5
	High	195.9	9.0	3.6	7.1	22.8	8.9	1.5	33.7	192.8	42.9	1.2	252.9	58.2	7.7
Suicidal Ideation	***Low***	189.0	8.5	3.1	7.3	21.6	7.1	2.1	28.1	231.4	47.2	1.3	263.0	46.6	9.0
	High	220.0	10.5	4.1	8.2	24.8	8.0	1.5	39.9	248.8	47.3	1.1	221.2	51.2	5.5

During the semantic fluency emotions task, a keyword analysis based on affect valence ratings demonstrated significant differences for distress *low* (0.46) versus *high* (0.40) severity groups, whereby the *high* severity group uttered keywords with lower valence. Unlike previous speech studies [14, 15] concerning depression, adolescents in the *high* severity group did not generate words with significantly lower valence than those in *low* severity group. For anxiety, depression, and suicidal ideation, the *low* and *high* severity groups produced similar average valence scores per recording (0.40 to 0.43). While these valence results do not align with previous free conversation speech studies [14, 15], which indicated that adults with mood disorders focus more negative terms, it should be noted the study herein involved adolescents and non-conversational speech samples.

In Table 3, semantic fluency food category nutrimental attribute keyword results demonstrated many key differences between *low* and *high* severity groups. For example, the distress *high* severity group produced significantly lower fibre (1.6g) and calcium (43.8mg) when compared to the *low* severity distress group (2.4g, 52.3mg). For depression, the *high* severity group produced the highest keyword averages for sugar (8.9g) and lowest for sodium (192.8mg) when compared to other disorders. Interestingly, high sugar consumption is linked to higher depression prevalence [20]. Moreover, the *high* severity suicidal ideation group had seven nutrimental attributes that were significantly different from the *low* severity group. Participants in the suicidal ideation *high* severity group produced keywords with significantly reduced vitamin C and potassium values when compared to the *low* severity group (i.e., these values were also much lower when compared to other disorder *high* severity values).

For all four mental illness self-surveys, the *high* severity groups, produced food keywords lower in fibre, iron, and vitamin C. Food keywords low in sodium and high in sugar were a more unique indicator for depression, whereas food keywords low in calcium were more a unique indicator for distress. Interestingly, for the four mental illness self-surveys, Table 3 results show that individuals with *high* severities uttered keywords significantly lower fibre averages than the low severity groups. Additional food keyword metrics, such as food subtypes (e.g., dairy, meat, ultra-processed), metabolomics (e.g., B-vitamins, antioxidants) and/or costs might also be useful during food semantic fluency tasks to determining elevated mood disorder severity associations.

During the sports semantic fluency task, sport metric results indicated that the depression and suicidal ideation *high* severity groups demonstrated an increase in sport keywords that were team-based (0.26, 0.36) with direct physical opponent interaction (0.70, 0.73) than the *low* severity groups who showed more balanced solo/team-based (0.11, 0.07) and opponent interaction (0.57, 0.54). For the sport location type metric, *low* and *high* severity group ranges were similar across the different mood disorders (0.42 to 0.49), whereby sports in outdoor environments were more frequently verbalized. Further analysis of the sports keyword analysis demonstrated that the number of on-field teammates was significantly higher for depression and suicidal ideation *high* severity groups (6.51, 6.85) when compared to the *low* severity groups (5.71, 5.60).

5. Conclusion

This study demonstrates that deeper analysis of the keywords spoken during semantic fluency categorical tasks, especially concerning habitual topics, can provide insights into individuals' mental lexicons. With regards to verbal semantic fluency tasks, it is shown that in addition to traditional semantic fluency correctness type metrics, new category-specific quantitative keyword indices, such as food nutritional and color shade metrics, can help identify mental lexical biases associated with increased mood disorder symptom severities.

6. Future Work

Further investigation into semantic fluency topic categories (e.g., beverages, fears, television programs) and topic-specific indices should be explored to discover additional biased lexical behaviors associated with specific mood disorders. Experimentation including non-English language and/or bilingual speakers could help to uncover whether word retrieval bias is universally exhibited across different languages/cultures. Other illnesses (e.g., autoimmune, neurological) that impact language skills could be evaluated using semantic fluency tests with topic-specific index metrics. This may help to find unique categorical word retrieval behaviors in certain illnesses; and further, possibly provide new metrics for monitoring patients' performance improvement during therapies. Due to the isolated, word-level production during recorded verbal semantic fluency tasks, automatic speech recognition with topic-specific keyword index text-processing analytics is a practical future consideration.

7. Acknowledgements

The Future Proofing Study funded by an NHMRC Project Grant (GNT1138405). We thank the Black Dog Institute (NSW, Australia), A/Prof. Aliza Werner-Seidler and Prof. Helen Christensen for their research dataset project support.

8. References

[1] The Centers for Disease Control and Prevention (CDC), "Youth Risk Behavior Survey Data Summary and Trends Report: 2011-2021", Official Report, 2021.

[2] Patel, V., Flisher, A.J., Hetrick, S., and McGorry, P., "Mental health of young people: a global public-health challenge", *The Lancet*, vol. 369 (9569), pp. 1302–1313, 2017.

[3] Copeland, W.E., Wolke, D., Shanahan, L., and Costello, J., "Adult functional outcomes of common childhood psychiatric problems: a prospective, longitudinal study", *JAMA Psychiatry*, vol. 72 (9), pp. 892–899, 2015.

[4] World Health Organization (WHO), "Adolescent & young adult health", 2023, downloaded: https://www.who.int/news-room/fact-sheets/detail/adolescents-health-risks-and-solutions

[5] Liu, L., Villavicencio, F., Yeung, D., Perin, J., and Lopez, G., "National, regional, and global causes of mortality in 5–19-year-olds from 2000 to 2019: a systematic analysis", *The Lancet Global Health*, vol. 10 (3), pp. e337–e347.

[6] Nezu, A.M., Nezu, C.M., Friedman, J., and Lee, M., "Assessment of depression", In: C.L. Hammen, I.H. Gotlib, Handbook of Depression (2nd Eds.), The Guilford Press, New York - USA, pp. 45–68.

[7] Sanchez-Moreno, J., Martinez-Aran, A., Tabares-Seisdedos, R., Torrent, C., Vieta, E., and Ayuso-Mateos, J. L., Functioning and disability in bipolar disorder: An extensive review. *Psychotherapy and Psychosomatics*, vol. 78 (5), pp. 285–297, 2009.

[8] Klumpp, H. and Deldin, P., Review of brain functioning in depression for semantic processing and verbal fluency, *Intern. J. of Psychophysiology*, vol. 75 (2), pp. 77–85, 2010.

[9] Gawda, B., and Szepietowska, E., "Trait anxiety modulates brain activity during performance of verbal fluency tasks", *Frontiers in Behavioral Neuroscience*, vol. 10 (10), pp. 1–15, 2016.

[10] Stasak, B., Joachim, D., and Epps, J., "Breaking age barriers with automatic voice-based depression detection", In: *Proc. IEEE Pervasive Computing*, vol. 21 (2), pp. 10–19, 2022.

[11] Joorman, J. and Gotlib, I.H., "Emotion recognition in depression: relation to cognitive inhibition, *Cog. & Emotion*, vol. 24 (2), pp. 281–298, 2010.

[12] Weintraub, M., Posta, F., Ichinose, M.C., Arevian, A.C., and Miklowitz, D.J., "Word usage in spontaneous speech as a predictor of depressive symptoms among youth at high risk for mood disorders", *J. Affect Disord.*, vol. 323, pp. 675–678, 2023.

[13] Zimmermann, J., Brockmeyer, T., Humm, M., Schauenberg, H., and Wolf, M., "First-person pronoun use in spoken language as a predictor of future depressive symptoms: preliminary evidence from a clinical sample of depressed patients", *Clinical Psychology & Psychotherapy*, vol. 24 (2), pp. 384–391, 2016.

[14] Gumus, M., DeSouza, D.D., Xu, M., Fidalgo, C., Simpson, W., and Robin, J., "Evaluating the utility of daily speech assessments for monitoring depression symptoms, *Digital Health*, vol 9, pp. 1–11, 2023.

[15] Yang, C., Zhang, X., Chen, Y., Li, Y., Yu, S., Zhao, B., Wang, T., and Luo, L., "Emotion-dependent language featuring depression", *J. of Behavior Therapy and Experimental Psych.*, vol. 81, 2023.

[16] Jacka, F.N., Rothon, C., Taylor, S., Berk, M., and Stansfeld, S.A., "Diet quality and mental health problems in adolescents from East London: a prospective study", *Social Psychiatry and Psychiatric Epidemiology*, vol. 48 (8), pp. 1297–1306, 2012.

[17] Khalid, S., Williams, C.M., and Reynolds, S.A., "Is there an association between diet and depression in children and adolescents? A systematic review", *British J. of Nutrition*, vol. 116 (12), pp. 2097–2108, 2017.

[18] Saghafian, F., Hajishafiee, M., Rouhani, P., and Saneei, P., "Dietary fiber intake, depression, and anxiety: a systematic review and meta-analysis of epidemiological studies", *Nutritional Neuroscience*, vol. 26 (2), pp. 108–126, 2021.

[19] Ding, J. and Zhang, Y., "Associations of dietary vitamin C and E intake with depression: a meta-analysis of observational studies", *Frontiers Nutrition*, vol. 9, pp. 1–16, 2022.

[20] Knüppel, A., Shipley, M.J., Llewellyn, C.H. et al., "Sugar intake from sweet food and beverages, common mental disorder and depression: prospective findings from the Whitehall II study", *Scientific Reports*, vol 7, pp. 1–10, 2017.

[21] Sutton, T.M, and Altarriba J., "Color associations to emotion and emotion-laden words: A collection of norms for stimulus construction and selection", *Behav. Res. Methods*, vol. 48 (2), pp. 686–728, 2016.

[22] Meier, B., Robinson, M.D., and Clore, G.L., "Why good guys wear white: automatic inferences about stimulus valence based on brightness", *Psych. Science*, vol. 15 (2), pp. 82–87, 2004.

[23] Pluhar, E., McCracken, C., Griffith, K.L., Christino, M.A., Sugimoto, D., and Meehan III, W.P., "Team sport athletes may be less likely to suffer anxiety or depression than individual sport athletes", *J. Sports Sci. Med.*, vol 18 (3), pp. 490–496, 2019.

[24] Werner-Seidler, A., Huckvale, K., Larsen, M.E. et al., "A trial protocol for the effectiveness of digital interventions for preventing depression in adolescents: the future proofing study", *Trials*, vol. 21 (1), pp.1–21, 2020.

[25] O'Dea, B., Han, J., Batterham, P.J., Achilles, M.R., Calear, A.L., Werner-Seilder, A., Parker, B., Shand, F., and Christensen, H., "A randomized controlled trial of a relationship-focussed mobile phone application for improving adolescents' mental health", *J. of Child Psych. and Psych.*, vol. 61 (8), pp. 899–913, 2020.

[26] Süsstrunk, S., Buckley, R., and Swen, S., "Standard RGB color spaces", In: *Proc. IS&T/SID 7th Color Imaging Conf.*, vol. 7, pp. 127–134, 1999.

[27] Mohammad, S.M., "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words", In: *Proc. 56th Annual Meeting of the Assoc. for Comp. Ling.*, Melbourne, VIC - Australia, pp. 174–184, 2018.

[28] United State Department of Agriculture (USDA), *Agricultural Research Service*, downloaded Sept. 2023: https://fdc.nal.usda.gov

[29] Kim, T.K., "T test as a parametric statistic", *Korean J. Anesth.*, vol. 68 (6), pp. 540–546, 2015.

[30] Hedges, L.V., and Olkin, I., Statistical Methods for Meta-Analysis, Academic Press, Orlando, FL - USA, 1985.

[31] Kelly, K. and Preacher, K., "On effect size", *Psych. Methods*, vol. 17 (2), pp. 137–152, 2012.

Towards Identifying the Common Social Emotions in Spoken te reo Māori: A Community-Oriented Approach

Himashi Rathnayake[1], Jesin James[1], Ake Nicholas[1], Gianna Leoni[2], Catherine I. Watson[1], Peter Keegan[1]

[1]The University of Auckland, New Zealand, [2]Te Hiku Media, New Zealand

hrat069@aucklanduni.ac.nz

Abstract

Emotion-based studies, including speech emotion recognition, often prioritise well-resourced languages, overlooking low-resource and Indigenous languages like New Zealand's te reo Māori. In these languages, emotions are not even well-defined. This study aims to identify common social emotions in te reo Māori speech. We designed a questionnaire using te reo Māori media recordings and collected feedback from te reo Māori speakers on the emotions. Our analysis yielded 218 emotion terms in te reo Māori and English. The study highlights the need for further categorisation of these emotion terms based on similarity using feedback from experts in te reo Māori.

Index Terms: emotions, speech technology, low-resource, Indigenous

1. Introduction

Identifying emotions from speech, known as speech emotion recognition (SER), is important in human-computer interaction (HCI) due to the demand for contactless interaction. However, such speech technology development is happening only in 100 out of the over 7,000 world languages, primarily in a few well-resourced languages like American English [1]. Very limited work on SER exists in low-resource and Indigenous languages. For instance, there is no SER system for te reo Māori (the Māori language). This study aims to establish the common social emotions in te reo Māori, which is a prerequisite for any emotion-based study in te reo Māori, including SER.

1.1. Te reo Māori

Te reo Māori is the only official Indigenous language of Aotearoa New Zealand. It is a Polynesian language fluently spoken by 71,000 people, which is 1.5% of New Zealand's population [2]. From 1847 until the 1960s, colonial practices actively discouraged the use of te reo Māori, promoting the use of English instead. These efforts resulted in a significant decline in te reo Māori-speaking community. Even today, only 7.9% of the Māori population are fluent speakers of te reo Māori [3], while most Māori remain bilingual, speaking both English and te reo Māori [2]. In the 1980s, the revitalisation of te reo Māori began. As a result, te reo Māori gained its official status by the Māori Language Act 1987. The government also started supporting te reo Māori in various areas. Te reo Māori education was supported by establishing te reo Māori schools and tertiary institutions [4]. Moreover, organisations like Te Reo Irirangi o Te Hiku o te Ika (Te Hiku Media), a Māori community communication hub for radio, online TV and other media services, were established [5][1]. Such organisations have collections of te reo Māori speech data, which can be valuable resources for any study involving te reo Māori.

1.2. Emotions in cross-cultural contexts

Defining an emotion is challenging, and there is no universally agreed definition of an emotion [6]. However, it is generally agreed that emotions include cognitive processes, physiological adjustments, behavioural changes, and affective experiences such as arousal and pleasure [6, 7]. Another commonly discussed characteristic of emotions is their universality [8, 9]. Ekman's studies are well-known for verifying the universality of emotions between different cultures, including Brazil, the United States, Argentina, Chile, Japan, and the Indigenous population of Papua New Guinea [8]. In those studies, people were asked to match a story with facial expressions representing a few emotions. Due to the similarity in responses, Ekman concluded that emotions are universal and categorised emotions into six categories, which are defined in English: anger, disgust, fear, happiness, sadness, and surprise [8, 9].

Current emotion-based studies, including SER, classify emotions into the same set of basic emotions with a few additional categories (e.g., [10], [11], [12]) or their direct translations into the languages in which the technology is being developed (e.g., [13]). Several studies have identified limitations in the first approach, which relies on emotions defined in English, due to flaws in the design of Ekman's study. These flaws include the selection being forced on the set of facial expressions provided and the study being designed without any cultural knowledge or local community involvement [14, 15, 16]. There has been an ongoing debate that emotions can differ between languages due to cultural variation [14, 17, 18]. Hence, using emotions defined for English may not adequately represent emotions in all languages. Additionally, these studies were based on facial expressions, and there has been no verification that the same emotions exist in speech. The second approach, which uses the translations of emotions in target languages, also has limitations when applied to Indigenous languages like te reo Māori. A study has found that Māori and Pākehā (European) have different patterns of emotions for different social situations [19]. Scholars have even suggested that nuances of Māori emotions can be lost when referenced in English [20]. Further, te reo Māori contains a wide vocabulary of emotions which do not have a direct English translation [21]. Therefore, using emo-

[1] Te Hiku media has been leading the preservation of te reo Māori through artificial intelligence (https://time.com/collection/time100-ai-2024) and leading the principles of data sovereignty for Indigenous data.

tions defined in English or the closest translations of English emotion terms might be inadequate for te reo Māori. Thus, it is crucial to identify the common social emotions in te reo Māori before proceeding with any technology development.

Community engagement is crucial in any study focusing on Indigenous languages. The studies developed without community engagement fail to represent the Indigenous perspectives accurately and are biased against Indigenous communities [22]. Hence, this study was conducted in close collaboration with the community to ensure it benefits the community. Further details on community collaboration will be presented in Section 3.5.

To address the need for emotion identification for te reo Māori through a community-oriented approach, this study aims to identify the common social emotions in te reo Māori speech through feedback from te reo Māori speakers. The results of this study will indicate whether Ekman's emotion model effectively captures te reo Māori emotions or if there is a need for additional or alternative emotion categories. This study will be the first research of its kind into Indigenous emotions-based studies.

2. Methodology

This section outlines the methodology followed in identifying common social emotions in spoken te reo Māori. Given the lack of prior studies with a similar focus, we developed a novel methodology for this study based on an online questionnaire. We used the questionnaire to collect feedback from te reo Māori speakers who know how emotions are categorised in their culture. We gave them different te reo Māori speech recordings as stimuli and asked them to comment on the emotions in the recordings.

2.1. Emotional speech collection

Since there are no emotional speech databases for te reo Māori, we used 100 te reo Māori recordings from the University of Auckland (UoA) library (50 hours of data) [23] and another 120 recordings (20 hours of data) shared by Te Hiku Media for research purposes [5].

UoA library resources include te reo Māori speech from various te reo Māori TV shows, as shown in Figure 1. Here, the Spongebob and Penguins of Madagascar versions were te reo Māori dubbed versions. The various topics included in Te Hiku Media recordings are shown in Figure 2.

The raw recordings from both sources contained some time durations with no notable emotions and some with only music and no speech. Hence, two research assistants who were schooled in New Zealand listened to the recordings and marked the durations with any notable emotion. These durations will be referred to as audio clips in this paper. They marked 745 audio clips (8 hours) to have some notable emotions. Since getting feedback for all 745 audio clips from te reo Māori speakers is not practical, we randomly picked 100 audio clips. We trimmed each audio clip to be 10-30 seconds in duration. Two of the authors, who are both Māori and fluent te reo Māori speakers, further checked and adjusted the audio clips to ensure that no clip ends in the middle of a phrase.

2.2. Questionnaire design

We then designed the questionnaire to obtain the feedback of te reo Māori speakers regarding emotions conveyed in te reo Māori speech. The test was bilingual, provided in both English and te reo Māori, allowing participants to choose their preferred language. We decided to make the questionnaire bilingual mainly due to te reo Māori being a revitalised language; even native te reo Māori speakers might have difficulty in coming up with te reo Māori emotion vocabulary. Moreover, Māori people may use English to express emotions while still conveying an emotional framework that is culturally Māori. As the literature suggests, English emotion terms might not be sufficient to describe Māori emotions; hence, the questionnaire will try to collect emotions conveyed in te reo Māori speech regardless of the language. The questionnaire consisted of the following questions.

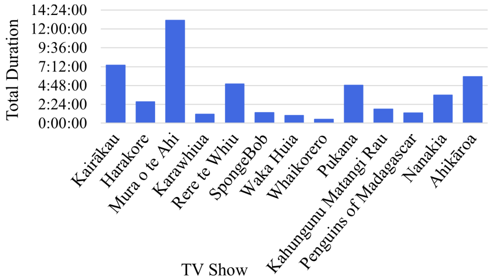


Figure 1: *UoA recordings - duration vs TV show*

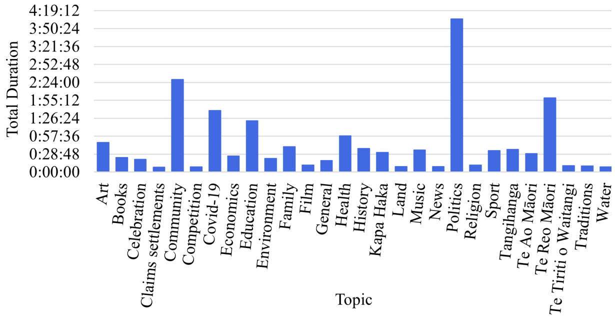


Figure 2: *Te Hiku Media recordings - duration vs topic*

1. Demographic questions - First, we asked for demographic information to understand the participants' background, including age, gender, Māori language ability, exposure to the Māori language, and hearing difficulties.

2. Emotion perception questions - We divided the selected 100 audio clips into five sets, assigning one randomly chosen set to each participant. We then asked the participants to listen to the audio clips and write down any emotions they could identify. The participants could respond in English or te reo Māori; the answers were free responses and could be descriptive.

3. Feedback questions - Finally, we asked the participants to write down the speech features that helped them identify the emotions in the given audio clips and the difficulties they faced in identifying emotions.

2.3. Participant recruitment and ethical considerations

The Māori researchers in the team advertised the questionnaire via email and social media to potential participants. Any fluent te reo Māori speaker with normal hearing ability and over 18 years old could participate in the questionnaire.

We obtained the ethics approval (Reference No. 25668) for the study from the UoA Human Participants Ethics Committee and obtained informed consent from all participants to ensure that ethical standards and cultural appropriateness are maintained.

2.4. Emotion terms filtering

Since the emotion perception questions allowed free responses, participants could provide more descriptive answers than a list of emotion terms. Therefore, the first author manually marked the key terms that describe the emotions in all the responses. These markings were then cross-checked by the Māori researchers involved in the study. Further, we converted all the terms into their adjective form and removed duplicates. Then, we removed the intensity words which describe emotions. For example, we removed terms like *very* and *āhua (somewhat)* if used with emotions (e.g., very angry, āhua pukuriri (somewhat angry)). Some terms from the questionnaire responses described actions (what the speakers of the recordings were doing) or the qualities of speakers instead of internal emotions. Hence, we removed such terms. For example, we removed terms such as blaming, laughing, akiaki (encouraging someone else), and whaikōrero (formal speech).

3. Results and discussion

3.1. Participant demographic information

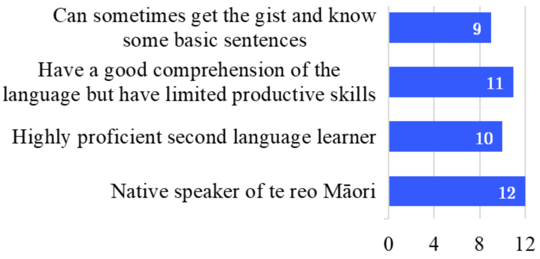


Figure 3: *Māori language ability of the participants.*

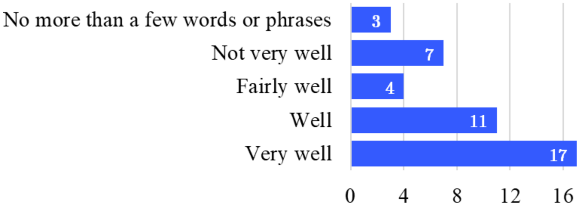


Figure 4: *Ability of the participants to speak and understand spoken te reo Māori in day-to-day conversation.*

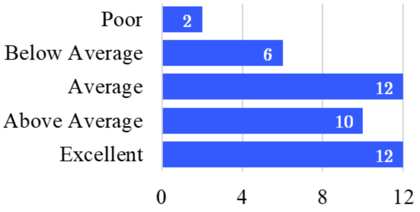


Figure 5: *Self-rated confidence of identifying emotions in spoken te reo Māori.*

Table 1. *Examples of responses received for emotional questions of the questionnaire (Te reo Māori emotion terms are marked in red, and English ones are marked in blue.)*

Response	English Translation
1. sadness, pōuri, crying, anger, riri, sorrow	sadness, sad, crying, anger, angry, sorrow
2. Hōhā ana te tāne. Tau ana te wahine.	The man is annoyed. The woman is calm.
3. Started excited / encouraging leading up to the bit, ka tahi ka hōhā te kaikōrero o tērā bit	Started excited / encouraging leading up to the bit, and then the speaker got annoyed
4. He pukuriri he hōhā hoki nō te wahine nei i te kore mauranga atu o ngā tamariki kia werohia hei ārai mate, māuiui anō hoki.	The woman is angry and annoyed that the children aren't being taken to get vaccines to protect them from death and illness.

Forty-two participants completed the questionnaire from late 2023 to early 2024. Among them, 28 identified as female, 12 as male, and two identified as gender diverse. The participants were between the ages of 22 and 73. One participant reported having inflamed eardrums.

Thirty participants reported English as their first language, whereas 10 participants reported their first language as te reo Māori. The other two participants reported English and te reo Māori as their first language. As Māori are bilingual speakers of English, this language distribution is expected in New Zealand.

The participants described their Māori language ability as shown in Figure 3. The figure shows that 12 participants identified themselves as native te reo Māori speakers, and 10 identified as highly proficient te reo Māori speakers. Further, participants rated their ability to understand spoken te reo Māori in day-to-day conversation, as shown in Figure 4.

Figure 5 shows participants' self-reported confidence in identifying emotions in spoken te reo Māori. Overall, the majority of the participants demonstrated significant confidence in identifying emotions.

3.2. Emotion terms identified from questionnaire responses

A few examples of responses received for emotion questions of the questionnaire are given in Table 1. The table provides the closest English translation of each response, which may not fully capture the nuances of the Māori emotions but is given to provide a general sense of the examples.

After removing duplicates and non-emotional terms, as detailed in Section 2.4, a total of 218 terms remained: 105 in te reo Māori and 113 in English. Figure 6 gives a word cloud of remaining emotion terms. The font size of each emotion term represents its frequency of appearance within the questionnaire responses. The word cloud illustrates the diversity of emotion terms in each language. In particular, we can find various emotion terms in te reo Māori that correspond to the same English emotion. For instance, terms like harikoa, hari, koa, koakoa, and hākoakoa all refer to the happy emotion. However, this prompts the question of whether these terms are exactly similar.

3.3. Participant feedback

The participants mentioned various factors that helped them identify emotions from the given speech, such as the speed,

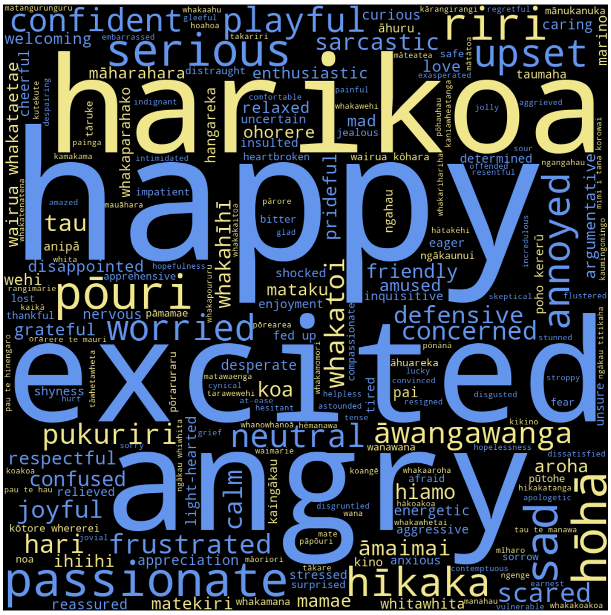


Figure 6: *A word cloud of emotion terms identified from questionnaire responses*

tone of the voice, volume, background noises, laughter, context, pitch, wording, intention, intonation, syllable stress, pronunciation, and kiwaha (colloquial sayings). Further, some participants mentioned that they just put themselves into the speaker's situation in the audio clips and imagined their feelings.

The participants mentioned the following as the difficulties they faced while identifying the emotions in the audio clips provided in the questionnaire. They mentioned that sometimes, they did not know the correct terms for some emotions. Moreover, the speakers' fast talk, monotone, talking over each other, background music, short audio clips, and lack of context in some audio clips made the task difficult. Further, a few participants mentioned that the task was hard because they could not see facial expressions. However, the study aims to find out the emotion in te reo Māori speech; hence, emotions identified through facial expressions were not expected.

3.4. Discussion

Let us consider the ten frequently found emotion terms within the participant responses: **happy**, **angry**, excited, **harikoa (happy)**, passionate, hōhā (annoyed/ boring), **pōuri (sad)**, **sad**, **riri (angry)**, hīkaka (eager/ excited). For a moment, let us set aside the issue of nuances being lost in te reo Māori to English translation. If te reo Māori terms are converted to the closest English translation and duplicates are removed, the remaining terms will be happy, angry, excited, passionate, annoyed, and sad. Three of those terms, happy, angry and sad, belong to Ekman's six basic emotions, whereas others are not explicitly listed in Ekman's emotions [9]. On the other hand, Ekman's remaining emotions were also present in the participant responses but were not as frequent. These findings suggest that Ekman's emotions may not fully capture common social emotions in te reo Māori, aligning with the discussion in Section 1.2. However, the frequency of emotion terms could vary depending on the selected audio clips; hence, further analysis of emotion terms is needed to finalise the emotion categories.

As mentioned in Section 3.2, different emotion terms found in te reo Māori related to the same English emotion within the responses. As highlighted in Section 1.2, the nuances can be lost when translating te reo Māori emotions into English. Hence, feedback from experts in te reo Māori is further needed to find whether such terms are exactly the same or if there are any nuances.

Due to the absence of an emotional speech database in te reo Māori, we used a Māori community media collection. Even though it may not capture all relevant emotions, community media could be a promising first step for developing emotion studies in any Indigenous language lacking large databases.

As te reo Māori is a language that has undergone revitalisation, the participants might lack the emotional vocabulary, making the questionnaire challenging for some participants, as indicated in their feedback. Even in the language ability questions, some participants rated lower scores in their language ability in te reo Māori. However, we cannot judge their language ability solely based on their self-rating. Previous studies of bilinguals observed that they tend to overrate or underrate their language competence [24]. In particular, te reo Māori native speakers consistently underrate their language ability [25].

The questionnaire participants were from the general population and may have issues distinguishing emotions. The absence of precise emotional definitions also limits the study. Given that this study is the first to identify emotions in spoken te reo Māori, further studies could build on this work to verify and expand the findings.

3.5. Community-oriented work

The primary significance of this work is that the potential applications of identified social emotions across various fields, including linguistic studies, emotional well-being, or technology development, offer significant benefits for the community. Each step of this study was developed in close collaboration with the community. The team has prior experience in community-oriented work. Three of the authors are Māori researchers who provided continuous guidance to the team throughout the study and closely worked in recruiting the participants for the questionnaire. During the study, we especially focused on ensuring data sovereignty, which refers to Māori data being subject to Māori governance [26]. Any study outcome will be bound by the Kaitiakitanga Licence [27] to ensure Indigenous people's mana (authority) over data and intellectual properties.

4. Conclusion and future work

To address the issue of emotions not being well-defined in te reo Māori, this study designed a questionnaire to collect te reo Māori speaker feedback on emotions in speech. The study identified 218 emotion terms related to te reo Māori speech in English and te reo Māori. While some emotion terms seem to have a similar meaning, feedback from experts in te reo Māori is needed to conclude these overlaps.

The next step is to conduct a focus group with te reo Māori speakers to group similar emotion terms and find the final categories of te reo Māori emotions. The identified emotions will then be used to record the first te reo Māori emotional speech corpus and will be acoustically analysed to develop an SER system for te reo Māori. We believe this study will inspire researchers to culturally sensitively explore emotions in other Indigenous languages and offer a valuable model for their work.

5. Acknowledgements

The authors wish to acknowledge the Science for Technological Innovation National Science Challenge (SfTI) organisation and Te Hiku Media [5] for funding the study. The authors also wish to acknowledge the contributions of Matthew Prance and Abigail Croot from the University of Auckland for marking the durations with notable emotions. The authors also acknowledge the support and feedback from the Speech group at the University of Auckland throughout the project (https://speechresearch.auckland.ac.nz).

6. References

[1] James, J., Watson, C., and Gopinath, D. P., "Exploring text to speech synthesis in non-standard languages", in Australasian International Conference on Speech Science and Technology, Australasian Speech Science and Technology Association, 2016.

[2] Nicholson Consulting and Kōtātā Insight, He Ara Poutama mō te reo Māori, Wellington, New Zealand: Nicholson Consulting, 2021.

[3] James, J., Yogarajan, V., Shields, I., Watson, C. I., Keegan, P. J., Mahelona, K., and Jones, P.-L., "Language Models for Code-switch Detection of te reo Māori and English in a Low-resource Setting", in Findings of the Association for Computational Linguistics: NAACL 2022, 650–660, 2022.

[4] Reilly, M., Duncan, S., Leoni, G., and Paterson, L., Te Koparapara: An Introduction to the Maori World, Auckland University Press, 2018.

[5] Te Hiku Media. Online: https://tehiku.nz/, accessed on 08 Dec 2023.

[6] Cahour, B., "Emotions: Characteristics, emergence and circulation in interactional learning", in M. Baker, J. Andriessen, and S. Järvelä [Eds], Affective Learning Together, 52–70, 2013.

[7] Kleinginna Jr, P. R. and Kleinginna, A. M., "A categorized list of emotion definitions, with suggestions for a consensual definition", Motivation and Emotion, 5(4):345–379, 1981.

[8] Ekman, P. and Friesen, W. V., "Constants across cultures in the face and emotion", Journal of Personality and Social Psychology, 17(2):124–129, 1971.

[9] Ekman, P., "An argument for basic emotions", Cognition and Emotion, 6(3-4):169–200, 1992.

[10] Nam, Y. and Lee, C., "Chung-ang auditory database of korean emotional speech: A validated set of vocal expressions with different intensities", IEEE Access, 10:122745–122761, 2022.

[11] Muneer, V., Basheer, K., and Thandil, R., "Convolutional neural network-based automatic speech emotion recognition system for malayalam", Indian Journal of Science and Technology, 16(46):4410–4420, 2023.

[12] Mangalam, A. R., Singh, S., Lalremtluanga, C., Kumar, P., Das, R., Basu, J., and Chatterjee, S., "Emotion recognition from mizo speech: A signal processing approach", in IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics, (Ballari, India), 1–6, IEEE, 2022.

[13] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al., "A database of german emotional speech", in Interspeech, 1517–1520, 2005.

[14] Russell, J. A., "Culture and the categorization of emotions", Psychological Bulletin, 110(3):426–450, 1991.

[15] Russell, J. A., "Forced-choice response format in the study of facial expression", Motivation and Emotion, 17:41–51, 1993.

[16] Crawford, K., "Artificial intelligence is misreading human emotion", The Atlantic, 27, 2021.

[17] Shweder, R. A., Thinking through cultures: Expeditions in cultural psychology, Harvard University Press, 1991.

[18] Messaoudi, A., Haddad, H., Hmida, M. B., and Graiet, M., "Tuniser: Toward a tunisian speech emotion recognition system", in International Conference on Natural Language and Speech Processing, Association for Computational Linguistics, 2022.

[19] Banks, C., A comparison study of Maori and Pakeha emotional reactions to social situations that involve whakamaa, PhD thesis, Massey University, 1996.

[20] Elder, H., Aroha: Maori wisdom for a contented life lived in harmony with our planet, Random House, 2020.

[21] Ritchie, J., "Staying with the Troubles of Colonised Emotional Well-Being of Young Children in Aotearoa (New Zealand)", 2021.

[22] Jones, P.-L., Mahelona, K., Duncan, S., and Leoni, G., "Kia tangata whenua: Artificial intelligence that grows from the land and people", Ethical Space: International Journal of Communication Ethics, 2023(2/3), 2023.

[23] Te Tumu Herenga—Libraries and Learning Services, The University of Auckland. Online: https://www.auckland.ac.nz/en/library.html, accessed on 08 Dec 2023.

[24] Ascher, C., "Assessing bilingual students for placement and instruction", ERIC Clearinghouse on Urban Education, 1990.

[25] Potaka, L. and Cochrane, S., "Developing bilingual questionnaires: Experiences from new zealand in the development of the 2001 maori language survey", Journal of Official Statistics, 20(2):289, 2004.

[26] Te Mana Raraunga—Maori Data Sovereignty Principles. Online: https://www.temanararaunga.maori.nz, accessed on 08 Dec 2023.

[27] Te Hiku Media, "Whare Kōrero Kaitiakitanga License". Online: https://xn–wharekrero-v3b.nz/kaitiakitanga, accessed on 08 Dec 2023.

How does Prosody Distinguish Wh-question from Wh-declarative in Shanghai Chinese

Ling Bijun

Tongji University, La Trobe University
lingbijun@gmail.com

Abstract

Wh-words are ambiguous between interrogative and indefinite interpretations in Shanghai Chinese. This paper compares the prosody of wh-interrogative with wh-indefinite in string-identical wh-questions and wh-declaratives, to further our understanding of the interface between semantics and prosody. Results showed that the semantic distinction is realized through information packaging. The wh-interrogative is by default the focus with prosodic prominence (longer duration, larger f0 range, and higher f0), whereas the wh-indefinitive can never attract focus. Specifically, its following noun is focal and prosodically accented.

Index Terms: wh-word, Shanghai Chinese, prosody, focus

1. Introduction

It has been widely acknowledged that wh-words are ambiguous between interrogative and indefinite interpretations in many wh-in-situ languages (such as Korean: [1]; Japanese: [2]; Chinese: [3]). Take Korean for example, the wh-word "nwukwu" is ambiguous between an interrogative reading 'who' and an indefinite reading 'someone'. Thus, the sentence in (1) can be interpreted as: i) a wh-declarative with an indefinite reading, and ii) a wh-question with an interrogative reading [1].

(1) Yuna-ka nwukwu-lul mann-a
Yuna-NOM who -ACC meet-INT
i) Yuna is seeing someone. [wh-declarative]
ii) Who is Yuna seeing? [wh-question]

Therefore, a question is raised about how speakers differentiate this semantic ambiguity (indefinite vs interrogative) when they occur in syntactically identical sentences. Previous studies demonstrated that this semantic ambiguity can be differentiated by prosodic features such as pitch accent, boundary tones, and phonological phrasing.

For instance, Korean utilizes pitch accents and prosodic phrasing to differentiate wh-questions and declaratives containing wh-words: a wh-word bears a high-pitch accent in a wh-question, while it bears a low-pitch accent in wh-declaratives. Furthermore, accentual phrasings are deleted following the wh-interrogatives [1, 4], as illustrated in Figure 1. Similarly, in Japanese, a wh-word displays a sharp rise of f0 in a wh-question, followed by post-focal reduction which suppresses all lexical accents up to the end. At the end of the wh-question, an interrogative rise intonation is added, terminating post-focal reduction [5], as illustrated in Figure 2.

Different from Korean and Japanese, which are intonation language and pitch accent language respectively, Mandarin Chinese is a tonal language with four lexical tones: High, Rise, Low, and Falling [6]. Since the fundamental frequency in tonal languages carries the function of lexical differentiation in addition to indicating intonation, the study of its prosody is more complex. So far, there are only three studies investigating the prosody of wh-question and wh-declarative in Mandarin Chinese: [3], [7] and [8]. Results showed that wh-declaratives differ from wh-questions in terms of prosodic properties from the clause onset. (1) Wh-declaratives are longer than wh-questions starting from the subject and the pattern reverses at the wh-word. (2) Wh-declaratives are lower in f0 and smaller in f0 range than wh-questions at the wh-word and there is a f0 range compression in the post-wh-word region in wh-questions. (3) Wh-declaratives show a larger intensity range than wh-questions at the verb and the pattern reverses at the wh-word [3].

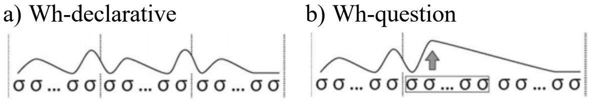

a) Wh-declarative
b) Wh-question
σ σ ... σ σ σ σ ... σ σ σ σ ... σ σ
σ σ ... σ σ σ σ ... σ σ σ σ ... σ σ

Figure 1: *The typical prosody pattern of a wh-declarative (a) and a wh-question (b) in Korean. The vertical dash line indicates the boundary of Accentual Phrases. The syllables surrounded by the box indicate the wh-phrase (cited from [1]).*

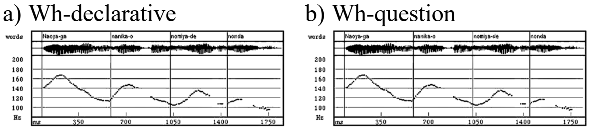


Figure 2: *The prosody pattern of the wh-declarative (a) "Naoya-ga nanika-o nomıya-de nonda." (Naoya drank something) and the wh-question (b) "Naoya-ga nani-o nomıya-de nonda no?"(What did Naoya drink). (cited from [5]).*

Overall, a general tendency can be summarized from these languages that wh-words are foci in wh-questions and are prosodically accented with raised f0 and expanded pitch range, but they cannot be foci in wh-declaratives as they are realized with compressed pitch contours. Furthermore, the post-wh-word region adaptations show language-specific features, such as the deletion of accentual phrases in Korean and the compression of the pitch range in Japanese and Mandarin Chinese.

To further our understanding of the interface between semantics and prosody, more cross-linguistic data is needed. Therefore, we take Shanghai Chinese as the object of study to shed some new light.

Shanghai Chinese, a Northern Wu dialect of Chinese, is spoken in the metropolis of Shanghai with a population of 20 million. In contrast to Mandarin Chinese, it has five citation tones [9] and its tonal system can be summarized by three sets of features (see Table 1): (1) Pitch register: high tones (T1, T2 and T4) with modal phonation and low tones (T3 and T5) with breathy phonation. (2) f0 contour: falling (T1) and rising (T2-T5). (3) Duration: Long tones (T1, T2 and T3), which occur in open or nasal-closed syllables [CV(N)], and short tones (T4 and T5), only occur in syllables closed by a glottal coda [CVʔ]. Furthermore, when syllables are combined into words, left-dominant tone sandhi rules are applied, which spreads the tonal contour of the initial syllable across the entire word [10, 11]. [12] proposed three types of prosodic units in Shanghai Chinese (prosodic word, prosodic phrase and intonational phrase) and tone sandhi only occurs within the prosodic word.

Table 1. *The value of citation tones and sandhi tones (using Chao's five-level numerical scale, which divides a speakers pitch range into five scales with 5 indicating the highest and 1 the lowest).*

Register	Tone type	Citation tone	Sandi tone (T+X)
High	T1 [HM]	53	55+31
	T2 [MH]	34	33+44
	T4 [Hq]	55	33+44
Low	T3 [LM]	13	22+44
	T5 [LMq]	12	11+13

All in all, the distinctive prosodic features of Shanghai Chinese make it an intriguing case for investigating the interface between semantics and prosody. Specifically, whether and how do speakers use prosodic cues to distinguish wh-questions from wh-declaratives in Shanghai Chinese when they occur in syntactically identical sentences?

2. Methodology

2.1. Stimuli

Each stimulus sentence consists of 11 syllables, forming 6 constituents: subject, VP, wh-phrase, object 1, preposition, and object 2, as illustrated in Table 2. Each sentence has both interrogative and indefinite interpretations. Take Stimuli ***IV*** for example, it could be interpreted as a wh-declarative "Old Wang bought several pieces of cake for Old Zhang." or as a wh-question "How many pieces of cake did Old Wang buy for Old Zhang?". For the speakers to produce sentences with correct semantics and intonation, each stimulus sentence is embedded in two dialogues. The wh-declarative is the answer to the question "What did Old Wang do?".

Table 2. *Stimulus sentences*

	Subject	*VP*	*Wh-phrase*	*Object 1*	*Prep*	*Object 2*
I	老王 ***/lɔ^{22} βã^{44}/*** ***Old Wang***	偷了 ***/thʊ^{55} ləʔ^{31}/*** ***stole***	几条 ***/tɕi^{33}diɔ^{44}/*** ***how many/several***	香烟 ***/ɕiã^{55} ie^{31}/*** ***Cigarette***	给 ***/pəʔ^{55}/*** ***for***	老张 ***/lɔ^{22} tsã^{44}/*** ***Old Zhang***
II	老王 ***/lɔ^{22} βã^{44}/*** ***Old Wang***	偷了 ***/thʊ^{55} ləʔ^{31}/*** ***stole***	几块 ***/tɕi^{33}kue^{44}/*** ***How many/several***	蛋糕 ***/de^{22}gɔ^{44}/*** ***cake***	给 ***/pəʔ^{55}/*** ***for***	老张 ***/lɔ^{22} tsã^{44}/*** ***Old Zhang***
III	老王 ***/lɔ^{22} βã^{44}/*** ***Old Wang***	买了 ***/ma^{22}ləʔ^{44}/*** ***bought***	几条 ***/tɕi^{33}diɔ^{44}/*** ***how many/several***	香烟 ***/ɕiã^{55} ie^{31}/*** ***Cigarette***	给 ***/pəʔ^{55}/*** ***for***	老张 ***/lɔ^{22} tsã^{44}/*** ***Old Zhang***
IV	老王 ***/lɔ^{22} βã^{44}/*** ***Old Wang***	买了 ***/ma^{22}ləʔ^{44}/*** ***bought***	几块 ***/tɕi^{33}kue^{44}/*** ***How many/several***	蛋糕 ***/de^{22}gɔ^{44}/*** ***cake***	给 ***/pəʔ^{55}/*** ***for***	老张 ***/lɔ^{22} tsã^{44}/*** ***Old Zhang***

2.2. Subjects and recording procedures

Five male and five female speakers, between the ages of 25 to 45, participated in the study. Both their parents and they were born and raised in urban areas of Shanghai. They were paid for their participation, and none reported any hearing, vision, or reading deficiencies. The recording was conducted in the sound booth at Tongji University. Each participant wore a Sennheiser PC 300 high-quality headset microphone that was positioned a constant distance (about 5 cm) from the speaker's mouth.

The sentences were presented in PowerPoint slides in a randomized order. Every participant was required to understand the meaning first and then read the material aloud in a natural way. Following a 5-minute break, the participants read the material again in a different random order. Therefore, we achieved 4 stimulus sentences * 2 context conditions * 10 speakers *2 times = 160 sentences.

2.3. Data analysis

The acoustic analysis was done in Praat [13]. Four layers were manually labeled (see Figure 3): (1) segment; (2) syllable; (3) prosodic word; (4) sentence. Then the Praat script "ProsodyPro" [14] was run, and it automatically provided us with (1) accurate f0 tracks by measuring f0 (Hz) at 10 equidistant points of the vowel; (2) max f0 (Hz), min f0 (Hz) and duration (ms) of each prosodic word; (3) max f0 (Hz), min f0 (Hz) and duration (ms) of the whole sentence.

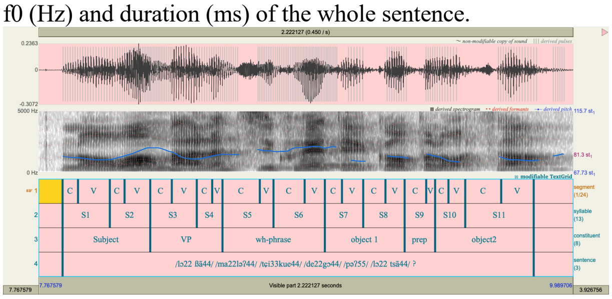


Figure 3: *Spectrogram with superimposed f0-contours of a male speaker.*

Subsequently, the f0 measurements in Hz were converted to semitone relative to 50 Hz using the formula in (1) to better reflect pitch perception [15]. Formula (1) relates frequency in semitones, F, to frequency in Hz, f:

$$F = 12*\log2(f/50) \tag{1}$$

Then f0 range of each prosodic word was calculated using the formula in (2):

$$\text{f0 range (st)} = \text{max f0 (st)} - \text{min f0 (st)} \tag{2}$$

To eliminate the significant individual differences in duration, the relative duration of each prosodic word was calculated by dividing the absolute value of duration by the average value of each speaker.

Linear mixed-effects models (using lme4 package) were used to investigate how duration and f0 measurements (f0 range, max f0, min f0) were affected by semantic interpretation and sentence intonation. All statistical analyses were carried out in R version 3.3.2 (R Core Team 2016) using the lme4 package version 1.1–12 [16].

3. Results

3.1. General description

Figure 4 displays the mean f0 contours of each stimulus sentence. These f0 contours were obtained by taking 10 f0 points (in Hz) at proportionally equal time intervals between the acoustic onset and offset of each vowel respectively. These values were then transformed into semitones and then averaged across speakers and repetitions according to sentence type. There are several things to be noted.

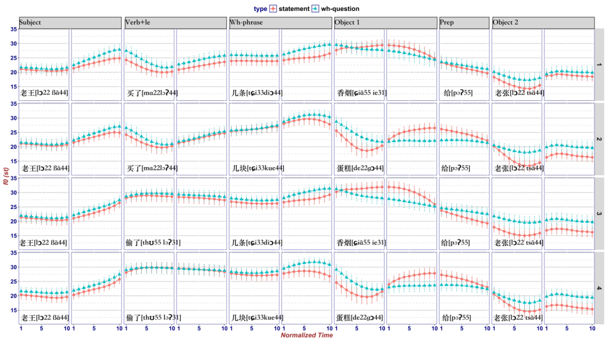


Figure 4: *The mean f0 contours of stimulus sentence. Red lines represent wh-declarative and blue lines represent wh-questions.*

The f0 contours of wh-questions are higher than wh-declaratives starting from the subject and the pattern reverses at object 1. Regarding the part from subject to wh-phrase, the biggest difference between the f0 contours of wh-questions and wh-declaratives occurs on the second syllable of wh-phrase (i.e. the measure word).

After the wh-phrase, the f0 curves of the wh-question begin to fall gently until the end of the sentence. In contrast, the f0 curves of the wh-declaratives show a prosodic accentuation at object 1, which is manifested as high tones higher and low tones lower. Therefore, concerning object 1 "/de^{22}gɔ^{44}/", the f0 curve of the first syllable in the wh-declarative is lower than that in the wh-question, while the curve of the second syllable is higher. Concerning object 1 "/ɕiã^{55} ie^{31}/", the f0 curves of both syllables in the wh-declarative are higher than their counterparts in the wh-question, with the maximum difference occurring on the second syllable. Such phenomenon indicates that object 1 is realized most prosodically prominent in wh-declaratives and it might be the foci.

After object 1, the f0 curves of the wh-declarative begin to fall sharply until the end of the sentence. Concerning object 2, its f0 curves in wh-declaratives are much lower than those in wh-questions, which indicates a PFC (post-focus compression) effect. The f0 contours of the preposition in wh-declarative interpolate between object 1 and object 2, as it is higher than that in wh-question following object 1 "/de^{22}gɔ^{44}/", while is lower following object 1 "/ɕiã^{55} ie^{31}/".

In sum, the wh-phrase is realized most prominently in wh-questions, while object 1 obtains the prosodical prominence in wh-declarative, which indicates the different focus allocations in two sentence types. Furthermore, the difference between wh-question and wh-declarative starts from the subject, which indicates that the distinction between questions and statements exists at the sentence planning stage.

3.2. Quantitative analyses

To verify the above observation, we first selected the relative duration, f0 range, max f0, and min f0 of the whole sentence as dependent variables. Linear mixed-effects models were conducted on these variables, respectively, with sentence type (wh-declarative vs wh-question) as a fixed factor and with the speaker as a random factor. Results are summarized in Table 3. The duration and f0 range of the wh-question are significantly smaller than those of the wh-declarative, which is in line with the results of Mandarin Chinese [3]. The min f0 of the wh-question is significantly higher, while there is no significant difference in the max f0.

Table 3. *The effects of sentence type on duration and f0 measurements of the whole sentence (with wh-declarative as the baseline).*

		Estimate	*S.E.*	*t*	*p*
duration	***(Intercept)***	1.061	0.014	77.725	0.000
	Wh-question	***-0.060***	***0.019***	***-3.102***	***0.002***
f0 range	***(Intercept)***	21.626	2.245	9.633	0.000
	Wh-question	***-4.456***	***1.105***	***-4.032***	***0.000***
max f0	***(Intercept)***	32.321	2.677	12.074	0.000
	Wh-question	-0.025	0.670	-0.037	0.970
min f0	***(Intercept)***	10.695	2.460	4.348	0.000
	Wh-question	***4.431***	***0.954***	***4.645***	***0.000***

Then the same Linear mixed-effects models were conducted on the relative duration, f0 range, max f0, and min f0 of each prosodic word respectively, with sentence type (wh-declarative vs wh-question) as a fixed factor and with speaker as a random factor.

Concerning the relative duration, the duration of wh-phrase in wh-question (Estimate=0.075, S.E.=0.036, t=2.064, p=0.039) is significantly longer than that in wh-declarative, while the duration of VP (Estimate=-0.050, S.E.=0.023, t=-2.190, p=0.029), object 1 (Estimate=-0.202, S.E.=0.028, t=-7.276, p<0.001) and preposition (Estimate=-0.077, S.E.=0.026, t=-2.997, p=0.003) in wh-question are all significantly shorter.

Concerning the f0 range, the f0 range of wh-phrase in wh-question is significantly larger than that in wh-declarative (Estimate=2.131, S.E.=0.592, t=3.599, p<0.001), while the f0 range of preposition (Estimate=-4.408, S.E.=0.691, t=-6.380, p<0.001) and object 2 (Estimate=-4.436, S.E.=1.032, t=-4.297, p<0.001) in wh-question are all significantly smaller, as illustrated in Figure 5.

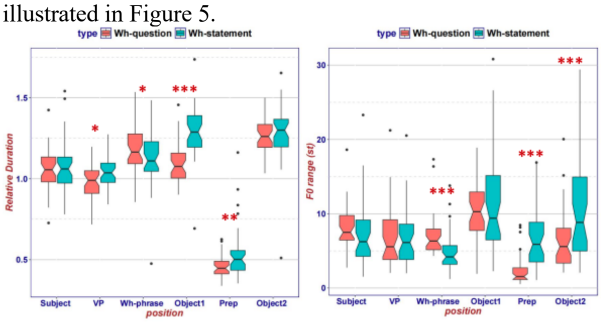


Figure 5: *The effects of sentence type on the relative duration (left) and f0 range (right) of each prosodic word.*

Concerning the max f0 and min f0, the max f0 of subject (Estimate=1.964, S.E.=0.455, t=4.320, p<0.001), VP (Estimate=1.457, S.E.=0.598, t=2.437, p=0.015) and wh-phrase (Estimate=2.695, S.E.=0.582, t=4.631, p<0.001) in wh-question are significantly higher than those in wh-declarative, while the max f0 of preposition (Estimate=-2.573, S.E.=0.654, t=-3.936, p<0.001) in wh-question is significantly lower. Meanwhile, the min f0 of preposition (Estimate=1.835, S.E.=0.910, t=2.016, p=0.044) and object 2 (Estimate=4.727, S.E.=1.080, t=4.375, p<0.001) in wh-question are significantly higher than those in wh-declarative, as illustrated in Figure 6.

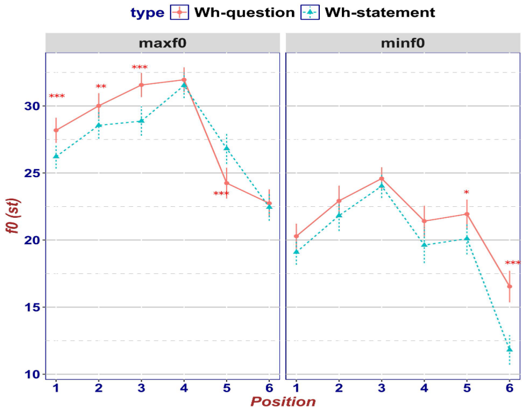


Figure 6: *The effects of sentence type on the max f0 (left) and min f0 (right) of each prosodic word. Red lines represent wh-declarative and blue lines represent wh-questions.*

4. Discussion

In this paper, we investigated the phonetic realization of wh-word in wh-questions (interrogative reading) and wh-declaratives (indefinite reading), to further our understanding of the interface between semantics and prosody. Our data demonstrated that speakers of Shanghai Chinese do use prosodic features (both global and local ones) to distinguish string-identical wh-declaratives and wh-questions.

The global features include the duration, f0 range, and max f0 of the whole sentence. Specifically, the duration of wh-questions is significantly shorter than that of wh-declaratives. Furthermore, the f0 range of the whole sentence in wh-questions is significantly smaller, which is mainly caused by the raised min f0 in wh-question, as there is no significant difference in the max f0 in the two sentence types, which is consistent with the intonation difference between questions and statements in Mandarin Chinese [17]. In other words, the global features reflect the contrast between interrogative and declarative intonation, rather than the prosodic difference of semantic ambiguation (interrogative vs indefinite).

We further examine the local features of each prosodic word. We found that the max f0 of wh-question is significantly higher than the wh-declarative starting from the subject and the pattern reverses at object 1. In contrast, there is no significant difference in the min f0 in the first half of the sentence until the preposition, where the min f0 of the preposition in the wh-question is significantly higher than that in the wh-declarative. The difference of min f0 between the two sentence types reaches the greatest at the final prosodic word, which indicates that the final prosodic word is the main unit carrying the interrogative intonation.

Furthermore, the local features also reflect the semantic ambiguity (interrogative vs indefinitive). In wh-questions, the wh-phrase is realized with significantly longer duration, larger f0 range, and higher max f0, which indicates that the wh-phrase is the most prosodically prominent word. In contrast, the wh-phrase is not accented in wh-declaratives, but its following noun (object 1) is prosodically accented, as evidenced by its significantly longer duration. Although we observed that the f0 contours of object 1 in wh-declarative are realized more distinctively in Figure 4 (i.e. high tones higher and low tones lower), there is no significant difference in f0 range, max f0, and min f0 between the two sentence types. How to explain such a phenomenon? There could be two possibilities. First, the wh-phrase and its following noun (object 1) form a prosodic phrase. As the wh-phrase is the focus of the wh-question, object 1 is also phonetically enhanced due to the carry-over effect, which caused no significant difference in f0 measurements between object 1 in the two sentence types. Second, it is difficult to prosodically enhance the object as proposed in [18]. Therefore, it relies more on duration adjustments rather than f0 adjustments.

In summary, prosody indeed differentiates wh-declarative from string identical wh-question in Shanghai Chinese. Generally speaking, wh-questions have a shorter duration, a smaller f0 range, and a higher min f0 than wh-declaratives. The max f0 of wh-questions is significantly higher than wh-declaratives starting from the subject and the pattern reverses at object 1. The min f0 of wh-questions is higher than wh-declaratives from the beginning and the difference keeps increasing until it reaches its maximum at the final prosodic word. Two implications can be derived from these results. Firstly, it can be observed that wh-declarative differs from wh-question in terms of prosodic properties from the sentence onset. This indicates that the distinction between question and declarative exists in the sentence planning stage. Secondly, the interrogative intonation is mainly reflected by min f0, especially on the final prosodic word. This is the main unit carrying the interrogative intonation.

Furthermore, the wh-word (interrogative) is the most prominent word in wh-questions, as it has a significantly longer duration, larger f0 range, and higher max f0. In contrast, the wh-word (indefinitive) is not accented in wh-declarative, instead, its following noun (object 1) is prosodically accentuated. Such results implicate the different focal status of wh-words in wh-questions and wh-declarative: wh-words are foci in wh-questions but cannot be foci in wh-declarative. In other words, semantics ambiguities (interrogative vs indefinitive) are realized through different information packaging, namely the allocation of focus. A questioned constituent (i.e. wh-interrogative in wh-question) is generally assumed to be the focus by default, whereas an indefinite pronoun (i.e. wh-indefinitive in wh-declarative) will never attract focus [19]. Such difference could be explained by information load. The questioned constituent carries the most important information in a question and therefore is the information focus. Instead, the indefinite pronoun modifies its following noun and its information load is relatively light, therefore its following noun becomes the information focus.

5. Conclusion

In this paper, we compared the phonetic realization of wh-questions to wh-declaratives, to further our understanding of the relation between semantics and prosody. We found that semantic interpretation is realized through information packaging (specifically focus allocation) and is represented by prosodic features. The wh-word (interrogative reading) in wh-questions is by default the focus with prosodic prominence (i.e. longer duration, larger f0 range, and higher max f0), while the wh-word (indefinite reading) can never attract focus in wh-declaratives. Instead, its following noun (i.e. object 1) is focal and prosodically accented. However, different semantic interpretations cause the string-identical sentences to be wh-question and wh-declarative respectively. It's difficult to say the above prosodic differences are caused by semantic or intonational distinction. Therefore, further investigation is needed, to further our understanding of the interface between semantics and prosody.

6. Acknowledgments

I owe my thanks to Professor Marija Tabain for very helpful comments on an earlier version of the paper. This work was supported by a grant from the Shanghai Office of Philosophy and Social Science (2022BYY003).

7. References

[1] Yun, J. “Meaning and prosody of wh-indeterminates in Korean,” Linguistic Inquiry. 50: 630–47, 2019.

[2] Kitagawa, Y. “When we fail to question in Japanese,” In Shinichiro Ishihara (ed.), Proceedings of the 2nd Workshop on Prosody, Syntax, and Information Structure (WPSI 2), Interdisciplinary Studies on Information Structure. University of Potsdam, Potsdam, 29-64, 2007.

[3] Yang, Y., Gryllia, S., and Cheng, L. L. “Wh-question or wh-declarative? Prosody makes the difference,” Speech Communication. 118: 21-32, 2020.

[4] Jun, S. A., and Oh, M. “A prosodic analysis of three types of wh-phrases in Korean,” Lang Speech. 39(1): 37-61, 2019.

[5] Ishihara, S. Intonation and Interface Conditions. PhD dissertation, MIT, 2003.

[6] Xu, Y. “Effects of tone and focus on the formation and alignment of F0 contours”. Journal of Phonetics, 27: 55–105, 1999.

[7] Dong, H. Issues in the semantics of Mandarin questions. PhD dissertation, Cornell University, 2009.

[8] Liu, X., Li, A., and Jia, Y. “How does prosody distinguish wh-statement from wh-question? A case study of standard Chinese,” In Proceedings of Speech Prosody, 1076-1080, 2016.

[9] Xu, B. H., and Tang, Z. Z. Shanghai Shiqu Fangyan Zhi (A Grammar of Inner-City Shanghai) (Shanghai Education Press, Shanghai, China), 1988.

[10] Zhu, X. N. Experimental Record of Shanghai Tones. (Shanghai Education Press, Shanghai), 2005.

[11] Ling, B. J., and Liang, J. “The nature of left- and right-dominant sandhi in Shanghai Chinese—Evidence from the effects of speech rate and focus conditions,” Lingua. 218: 38-53, 2019.

[12] Selkirk, E., and Shen, T. Prosodic domains in Shanghai Chinese. In S. Inkelas, & D. Zec (Eds.), The phonology-syntax connection. Chicago: University of Chicago Press, 313–337, 1990.

[13] Boersma, P. and Weenink, D. Praat: Doing phonetics by computer (Version 5.1.30) [Computer program]. Retrieved from http://www.praat.org, 2010.

[14] Xu, Y. “ProsodyPro — A tool for large-scale systematic prosody analysis,” In Proceedings of Tools and Resources for the Analysis of Speech Prosody. Aix-en-Provence, France, 2013.

[15] Rietveld, T., and Chen, A. J. “How to obtain and process perceptual judgements of intonational meaning,” In Sudhoff, S., et al. (ed.), Methods in empirical prosody research. Berlin: Walter de Gruyter, 283–319, 2006.

[16] Bates, D. Maechler, M., Bolker, B., and Walker, S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-12, 2014.

[17] Liu, F., and Xu, Y. “Parallel encoding of focus and interrogative meaning in Mandarin intonation,” Phonetica. 62: 70–87, 2005.

[18] Xu, Y. “Speech melody as articulatorily implemented communicative functions,” Speech Commun. 46: 220–251, 2005.

[19] Jackendoff, R. S. Semantic Interpretation in Generative Grammar. Cambridge, MA: MIT Press, 1972.

The Production of /l/ in Italian by HS and L2 Italian Speakers in Australia: Exploring the Effect of Syllable Position and Adjacent Vowel Quality

Valentina De Iacovo, John Hajek

Università degli Studi di Torino, The University of Melbourne

valentina.deiacovo@unito.it, j.hajek@unimelb.edu.au

Abstract

Heritage speakers (HS) are often noted for having a distinct sound or accent compared to both monolingual (L1) and second language (L2) speakers of the same language. In this study we focus on the production of the lateral approximant /l/ in Italian by HS and L2 speakers of the language in Melbourne (Australia) to investigate whether speaker status has an effect: in English both light (or clear) and dark (velarized) alveolar laterals occur depending on syllabic structure, while in Standard Italian /l/ is always a plain alveolar without secondary velarization. In this pilot study we conducted a production task with a set of words where /l/ occurred in different syllable positions and contexts, and which were read by three groups: HS, L2 and L1 Italian speakers. Results showed that the HS behave differently from both L1 and L2 speakers. However, while syllable position and structure influence the realization of /l/ between groups, there is no statistically significant difference within groups. Nevertheless, the quality of the adjacent stressed vowel influences the degree of lightness/darkness of the lateral sound across all groups.

Index Terms: bilingualism, lateral approximant, Italo-Australian community, heritage speakers, Italian L2

1. Introduction

1.1. Italian heritage speakers in Australia

Following WWII, more than 270,000 Italians moved permanently to Australia between 1947 and 1976 [1] – with the largest concentration establishing itself in Melbourne, the capital city of the State of Victoria. While those who migrated (so called *1st generation*) usually speak a regional dialect and/or a regional variety of Italian as L1 and Australian English (AuE) as L2, their children (*2nd generation*), born and educated in Australia, normally speak AuE as their L1 and/or a regional dialect and/or a regional variety of Italian as their heritage language. According to the most recent census conducted in 2021, more than 1.1 million Australian residents reported having Italian ancestry. Of these some 384,000 live in Victoria [2]. By far the largest proportion of Italian migrants and their descendants is from Central-Southern Italy (e.g. Sicily, Calabria and Abruzzo), although there was also significant permanent migration from Northeast Italy (specifically the Veneto and Friuli-Venezia Giulia regions).

A long tradition of linguistic research on the Italian community in Australia has to this point focused mainly on sociolinguistic practices and interactions within the community and on language shift [3, 4 for a review]. Some very recent research has now started to explore the potential effects of attrition and language contact from a phonetic perspective due to the bilingualism and/or trilingualism (regional dialect, regional Italian and AuE) of those speakers [5, 6]. With this contribution we aim to contribute further to this developing field of research by focusing specifically on the production of lateral /l/ in Italian. We compare separate groups of Italian HS and L2 speakers with a group of L1 Italian speakers to investigate the pronunciation of this consonant in a language contact context where at least some lateral velarization might be expected to appear in HS speech due to the effect of simultaneous proficiency in AuE.

1.2. Lateral /l/ in Italian vs Australian English

Traditionally, in Standard Italian (SI) the lateral approximant /l/ is reported to be fully alveolar without any secondary place of articulation [7]. However, there are also some very brief reports that in regional varieties of Italian it can also occur as semi-velarized or prevelarized depending on the syllabic context in different parts of northern (Veneto, Emilia-Romagna) and central-southern (Tuscany, Campania, Apulia, Sicily) Italy [8, 9]. It can also be velarized or retroflexed in some Italian dialects [10, 11] which then influences in turn the regional variety of spoken Italian. In terms of previous acoustic analysis of the lateral /l/, in Italian the only existing phonetic study we are aware of is based on a group of 10 Florentine Italian male speakers [12] where /l/ is analysed only in word-medial syllable-initial position within a CVˈCV sequence, e.g. *calata* 'dropped') and for which measured acoustic frequency ranges were reported to be 500-810 for F1 and 1165-1430 for F2.

In AuE, /l/ is usually described as surfacing as 'light' or 'clear' [l] in syllable-onset position, and velarised as [ɫ] i.e. 'dark' in coda position, while it is also often dark before a morpheme boundary preceding a vowel [13, 14]. However, [15] has suggested that syllable onset /l/ is more dorsal or velarized in AuE than in other varieties of English, claiming that the distinction clear vs dark is neutralized. This claim finds some confirmation in a recent study on AuE [16] in which initial electropalatographic (EPG) measurements for one speaker showed that lateral velarization appeared to occur equally in both onset and coda positions. That said, the general consensus still is the degree of velarization still appears to be relative according to syllable context in AuE: it is claimed specifically that velarization will always be greater in syllable-codas than in onsets, and in intervocalic word-medial ambisyllabic position the lateral may be intermediate with respect to relative velarization [13]. Such gradience in velarization related to syllable position has since been confirmed experimentally across different groups of AuE speakers [17] (see also below).

In addition to this, while the lateral is known also to influence adjacent vowel quality in AuE [18], the quality of vowels adjacent to the lateral consonant are also known to influence its relative darkness (e.g. [19]).

1.3. HS, L1, L2 and previous studies on lateral approximants in HS and L1 in migration settings

From a linguistic perspective, the term 'heritage speakers' (HS) sometimes also known as second generation migrants, refers to the children of the original migrants, who have lived in a bilingual/multilingual environment from an early age. While first generation immigrants are dominant in the native language of their home country, even if they may also have undergone L1 attrition, HS have as their dominant language, as noted above, the language of the host country [20]. From a phonetic perspective, there is also evidence that HS behave differently from L2 speakers who have acquired the L2 later in life. According to [21] for example, HS are able to maintain the Mandarin Chinese post-alveolar contrast in production and, in contrast to native L1 speakers and late L2 learners of Mandarin, tend to keep Mandarin and English sounds apart. [22] investigated the contrast between /l/ and /ɫ/ in L1 Albanian speakers resident in London and with L2 English to see what the impact, if any, of the allophonic distribution of lateral velarization in English might be. They found a stronger trend for light /l/ to become dark in coda position than for dark /ɫ/ to become light in onset position in Albanian. [17] investigated velarization of /l/ in the AuE of Anglo-Celtic and Lebanese Australians. As previously noted, results confirmed a positional effect, with /l/ darker word-finally than in word-initial position. Interestingly, the degree of darkness is also related to gender, ethnic identity, and social networks.

1.4. Aim of the study

If, on the one hand, SI has only a plain alveolar lateral, while AuE presents a clear or dark alveolar according to syllabic context, how do HS of Italian produce /l/ in Italian, particularly with respect to position within the syllable?

In this pilot study we investigate whether Italian HS and AuE L2 speakers produce /l/ in Italian as L1 speakers do and, if not, whether there is any difference between HS and L2. We also explore what if any potential differences are related to syllable structure and the preceding/following stressed vowel. Given the reported difference between SI and AuE with respect to the articulation of /l/, and the absence of any syllable effect in the former, and its reported conditioning effect in the latter, we therefore want to investigate: (1) if the place of articulation of the lateral /l/ produced by HS is 'clear' alveolar in all syllabic positions or, if as a result of HL attrition, and contact with and the dominance of AuE, /l/ is velarized, i.e. 'dark, in different syllable positions; (2) to see if velarization occurs to the same degree among HS and L2 speakers, given their shared proficiency in AuE; and (3) to see the extent to which the adjacent stressed vowel influences the degree of lightness/darkness of the lateral, assuming that back vowels (a, o, u) are more likely to contribute to lateral velarization than front vowels (i, e) [18]. In order to do so, we conducted a production experiment to determine the potential velarizing effect of specific syllabic structures on /l/ based also in part on the position in the word as well as the effect of the nature of the adjacent stressed vowel.

Our first hypothesis is that L1 Italian speakers will produce light alveolar /l/ in all contexts, while HS and L2 speakers will differentiate light and dark laterals based on onset or coda position respectively. Assuming that is the case, our second hypothesis is that /l/ produced by L2 will be darker than those produced by HS regardless of context – since they are likely to be the most influenced by their L1 AuE in contrast to HS who may be expected to gravitate towards L1 Italian norms given their longstanding exposure to and proficiency in Italian. As no previous acoustic studies on Italian have previously taken into account the role played by the syllabic structure and the potential influence of the adjacent stressed vowel, with this study we also aim at giving some first quantitative results about these matters more generally.

2. Experimental setting

2.1. Set of target words

In order to test potential differences in the realization of /l/ based on syllable structure, a set of words was created with the lateral in onset or coda position. Words employed (Table 1) were disyllabic (with initial stress) and for each category we always included 5 orthographic vowels (a, e, i, o, u). To do this, in a few cases, we also made use of nonsense words. We decided to test 3 possible syllabic structures where /l/ can occur in Italian:

- word-initial syllable-onset position followed by a stressed ('lVCV, *lato*)
- word-medial syllable-onset position preceded by a stressed vowel ('CVlV *pala*);
- word-medial syllable coda position preceded by a stressed vowel and followed by a voiceless stop /p, t, k/ ('CVlCV *salta*);

Table 1. *Target words used in the production task.*

'lVCV	lato	lega	lina	loro	Luca
'CVlV	pala	pela	fila	sola	mula
'CVlCV	talpa	scelta	milto	polpa	culpa
	salta	telca	pilco	volta	adulto
	palco	felpa	ilpa	solco	sulca

2.2. Production experiment

The experiment was created and hosted using the Gorilla platform [23] on a Dell Latitude 7490 laptop. Speakers read aloud once a list of written carrier phrases in Italian containing the target word such as "Ho detto *lato* proprio ora" (I said *side* just now) which were presented to them in random order. Recordings were made in a quiet room at the University of Melbourne, using a Rodelink Lav microphone. Only three tokens were not considered because of mispronunciation and a final set of 747 tokens was used for the purpose of analysis.

2.3. Participants

For this study we recruited 10 HS of Italian (henceforth HS), 10 L2 Italian speakers (EN) and 10 L1 Italian speakers (L1). The HS group consisted of women (aged 51-65), born in Melbourne or elsewhere in Victoria and with Italian background from the centre or south of Italy. They also speak Italian fluently, and habitually in the family. The EN group was made up of L1 English female students (aged 19-22), all Melbourne born and who were also studying Italian at post-beginner level at university. Their proficiency in Italian is education-derived. The L1 group involved L1 Italian speakers (aged 30-67) born and recorded in the north of Italy. All participants were naive to the purpose of the experiment and did not report any speaking or reading difficulties. In order to have a homogeneous group in terms of vocal tract, only female speakers were involved in this study.

2.4. Acoustic measurements

Target sounds were manually measured and labelled using Praat [24] according to the position of the lateral in three different syllable sequences (onset lateral /ˈlVCV/, /ˈCVlV/, and coda lateral /ˈCVlCV/) and the preceding or following stressed vowel (a, e, i, o, u). F1 and F2 formant values were extracted and downloaded into a csv file using a Praat script [25] which extracted 10 interval points of F1 and F2 measures for each target /l/ in the wav file and the associated Textgrid (see Results). Subsequently, the midpoints of the F1 and F2 estimated formants were automatically extracted and were also converted to the Bark scale using the formula recommended in [26]: F1/2 Bark = [(26.81 × F1/2) / (1960 + F1/2)] − 0.53. We present results here for mid-point only.

While F1 and F2 values on their own are important in understanding lateral articulations, the proximity between the two is considered to be the primary acoustic cue of relative velarization. Darker laterals have a lower F2, closer to F1, while clearer laterals have a generally lower F1 but a higher F2 [19]. The F2-F1 Bark measurement, used in previous studies to measure /l/ velarization [e.g. 27], was also adopted in order to better quantify the lateral's clearness or darkness from a perceptual perspective [17].

It is well-known that the acoustic identification of velarized versus vocalized laterals is difficult [19]. Since our analysis here relies exclusively on acoustic data, and given the pilot nature of our study including the relatively limited size of our sample, we have not tried through other means to identify and separate potentially vocalized tokens from the tested sample. Given the relatively low frequency (5.8%) of vocalized /l/ in previous experimental work on AuE [17], this may not be such an issue but remains nevertheless something for future non-acoustic assessment.

3. Results

3.1. F1 and F2 Hz and F2-F1 Bark

Our analysis was performed using R [27]. Mean values for lateral F1 and F2 at mid-point grouped by type of syllable structure and sequence (/ˈlVCV/, /ˈCVlV/, /ˈCVlCV/) and group (L1, HS, L2) are shown in Table 2. On initial inspection, it seems that variation occurs within the three groups according to syllable context. L1 group also shows more variation in F2 values compared to HS and L2. HS speakers show much lower F2 values in all the syllable contexts compared to the other groups, indicating that /l/ is generally darker (with the /ˈlVCV/ context clearer than for the other two tested contexts due to a noticeably lower F1). Based on F1 and F2 values, laterals in coda position are clearer for the L1 group and become progressively darker in turn for L2 and HS groups respectively. Unexpectedly, in both syllable onset positions (/ˈlVCV/ and /ˈCVlV/) L2 speakers show higher F2 values than L1 speakers, pointing to potentially clearer onset laterals for that group.

In Table 3 (with respective boxplots in Figure 1) we show the means and standard deviations of F2-F1 Bark normalization for the three syllable contexts investigated. We note firstly that there is variation in F2-F1 Bark within all three groups, with the L1 group having the highest average values (range 7.69-8.15) in all three categories when compared to the other two (HS 6.43-7.07, L2 7.42-7.78), corresponding to a generally more fronted place of articulation (i.e. more alveolar and less velarized). Particularly striking is the unexpectedly high 'clear' value (8.15) for L1 in syllable-final position. The HS and L2 groups on the other hand show gradient velarization according to syllable type and location: F2-F1 Bark values fall in a cline from word-initial to word-medial and then syllable-final position – consistent with previous reports for AuE ([13], [17]). The cline is, however, more moderate for L2 than it is for HS. At the same time overall F2-F1 Bark mean values for /l/ are much higher for both L1 and L2 groups than for HS (respectively 7.97 for L1, and 7.53 for L2 but only 6.63 for HS). Overall, the L1 group produces the clearest /l/ followed by L2 (especially in non-coda position), with much lower values for HS in all contexts indicating these speakers generally produce much more velarized laterals.

Table 2. *Lateral F1 and F2 midpoints' mean in Hz by type and group.*

Type	ˈlVCV		ˈCVlV		ˈCVlCV	
Group	F1	F2	F1	F2	F1	F2
L1	365	1595	408	1691	380	1743
HS	368	1459	414	1495	415	1446
L2	409	1732	452	1781	411	1661

Table 3. *F2-F1 mean in Bark and SD by type and group.*

Group		ˈlVCV	ˈCVlV	ˈCVlCV	Overall
L1	mean	7.71	7.69	8.15	7.97
	SD	1.42	1.57	1.77	1.67
HS	mean	7.07	6.78	6.43	6.63
	SD	1.53	1.83	2.28	2.07
L2	mean	7.78	7.61	7.42	7.53
	SD	1.81	1.77	2.29	2.10

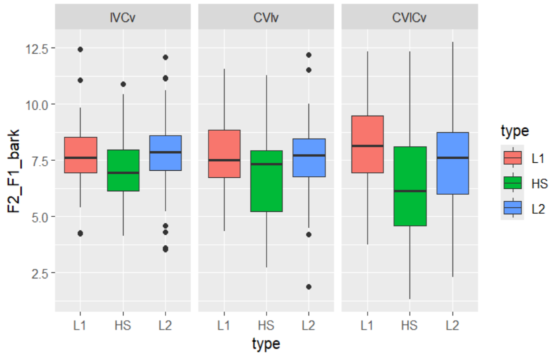


Figure 1: *Boxplots of F2-F1 Bark by type and group (L1, HS, L2).*

Before conducting any significance test, we first assessed the assumption of equality of variance with a Levene's test. Because the test showed that the assumption of equality was not met (p < .001), we decided to use a Games-Howell test which generally offers the best performance in these scenarios. The test revealed that F2-F1 Bark was significantly different overall among the three groups (HS-L1 and HS-L2 p values < .001, L1-L2 p < .05). As a next step, a sequence of the same test was conducted at syllable level to examine whether there were significant differences in F2-F1 Bark among the three groups. In coda position preceding a consonant (/ˈCVlCV/) the test revealed a significant difference between L1 v. HS and HS v. L2 (p < .001) as well as between L1 v. L2 (p < .05) for /l/. A significant difference was only found for /l/ in word-medial syllable-onset position following a stressed vowel (/ˈCVlV/)

between L1 and HS (*p* < .05), while no significant effect was found with respect to the word-initial syllable with lateral onset (/ˈlVCV/) amongst the three groups. Within group results were very different: surprisingly there was no effect of syllable context in any group - even in the case of coda /l/ v. word onset /l/ for HS (p = 0.66). This may be the result of variability seen in high standard deviation values.

3.2. Effect of the adjacent stressed vowel

Our third research question related to the extent to which the nature of the adjacent stressed vowel influences or correlates in some way with the degree of clearness/darkness of the lateral, assuming that back vowels would, due to tongue dorsum retraction, contribute more to velarizing the lateral compared to front ones. It should be borne in mind that the five vowels we tested are captured orthographically <a, e, i, o, u> and small differences in phonetic vowel quality between It and AuE are not unexpected, but were not explored here. Table 4, 5 and 6 show F2-F1 Bark mean values of the lateral midpoints for each of vowels in each of the three syllabic contexts, i.e. /ˈlVCV/, /ˈCVlV/ and /ˈCVlCV/ respectively. From a first overview, all groups (IT, HS, L2) show a difference in the degree of clearness/darkness on the basis of the front/back vowel distinction, regardless of syllable structure: back vowels show lower F2-F1 Bark differences, corresponding to a greater velarization of /l/. Comparing the three speaker groups, however, the lowest recorded values are for HS <a> and <o> before word-medial /l/ and syllable final coda /l/ in /ˈCVlCV/.

Based on those general results, we then conducted another Games-Howell test to test for statistical significance. In coda position (/ˈCVlCV/) results revealed a significant difference between L1 and HS when the lateral is preceded by <a>, <e>, <o>, <u> (*p* < .001) and between HS and L2 when it is preceded by <o> and <u> (*p* < .01). A significant difference is also found in onset position (ˈlVCv) between L1 and HS (*p* < .001) and between HS and L2 (*p* < .05) when /l/ is followed by <o>. Not surprisingly, within group comparisons also confirmed significant differences in vowel effects in the predicted direction, with significant difference in L1 and HS group (*p* < .001) between front and back vowels in coda position.

Table 4. *F2-F1 mean in Bark and SD by type and group.*

		ˈlVCV				
Group		_i	_e	_a	_o	_u
L1	mean	8.78	8.13	7.01	7.47	7.15
	SD	2.32	0.75	1.11	0.70	0.93
HS	mean	8.53	7.38	6.31	5.70	7.43
	SD	1.31	1.80	0.94	0.8	0.92
L2	mean	9.18	8.11	6.51	7.05	8.07
	SD	2.38	1.48	1.44	1.40	1.06

Table 5. *F2-F1 mean in Bark and SD by type and group.*

		ˈCVlV				
Group		i_	e_	a_	o_	u_
L1	mean	9.05	8.59	6.51	6.85	7.46
	SD	1.53	1.52	0.85	0.92	1.38
HS	mean	8.57	7.71	5.58	5.23	6.68
	SD	1.14	1.04	1.50	1.72	1.41
L2	mean	9.13	7.32	6.62	6.96	8.02
	SD	1.82	2.15	0.90	1.22	1.52

Table 6. *F2-F1 mean in Bark and SD by type and group.*

		ˈCVlCV				
Group		i_	e_	a_	o_	u_
L1	mean	9.80	9.15	6.92	6.98	7.80
	SD	1.24	1.13	1.37	1.25	1.65
HS	mean	8.62	7.07	4.99	5.32	6.13
	SD	2.40	2.26	1.10	1.72	1.68
L2	mean	8.77	7.61	6.14	6.71	7.89
	SD	2.98	1.54	1.80	1.82	2.19

4. General discussion and conclusion

In this contribution we have presented the results of a pilot study aimed at exploring the production of lateral approximant /l/ in HS and L2 speakers of Italian in Australia compared to L1 speakers. Overall, velarization varied across the three groups, with an effect most evident in coda /l/ and with none for word-initial /l/. The L1 group produced the clearest laterals overall – – consistent with descriptions of Standard Italian. That said, they also show some predictable vowel-conditioned velarization. The L2 group who could be expected to show significant velarization due to the influence of AuE also appeared to produce generally clear laterals. In coda position /l/ in both L2 and HS speakers is significantly different from IT, albeit much more velarized for HS than for L2. Overall, however, our results also show that L2 are able to produce laterals in a manner much more similar to L1 than HS, who instead tend to produce much more velarized laterals across the board. Neither of these findings was predicted. It is possible that the L2 group was hyperarticulating during the task, and in so doing avoiding velarization. We initially expected HS to land somewhere in the middle between L1 and L2 speakers, given the potentially counter-balancing influence of Italian and AuE. This was not the case – they had the darkest laterals in all syllable contexts. Moreover, a suggested syllable-conditioned cline in velarization gradience was not found to be significant: /l/ was relatively dark in all contexts. The reasons for overall darkness are not clear, but there are a number of possible explanations. It may, for instance, be due to the effect of a local Italo-Australian accent of Italian in which velarization is characteristic and borrowed from AuE. It may also lateral velarization is more widespread in Central and Southern Italy than is currently reported and has been retained in the speech of our HS speakers. Both of these hypotheses require further investigation.

With respect to vowel interactions, we predicted that for articulatory reasons of tongue retraction, laterals would show greater evidence of velarization when adjacent to the stressed back vowels (a, o, u). This was confirmed, with particularly greater effect for <o> for HS.

Future work, ideally based on a larger data sample, will explore the dynamics of formant trajectories to understand better the nature and process of velarization. It should also explore lateral production in the AuE of our two Australian groups, HS and L2, to see to what extent their results for Italian correlate with their articulation of /l/ in English.

5. Acknowledgements

The authors would like to thank the participants who took part in the task and Paul Scutti for his advice on statistics.

6. References

[1] Castles, S., “Italians in Australia: The Impact of a Recent Migration on the Culture and Society of a Postcolonial Nation”, Center for Migration Studies special issues, 11(3):342-367, 1994.

[2] Australian Bureau of Statistics, “Table 4 - Cultural diversity data summary”, Ancestry by State and Territory, accessed 11 June 2024, 2021.

[3] Rubino, A. and Bettoni, C., “Language maintenance and language shift: Dialect vs Italian among Italo-Australians”, Australian Review of Applied Linguistics, 21(1):21-39, 1998.

[4] Rubino, A., “Multilingualism in Australia: Reflections on current and future research trends”, Australian Review of Applied Linguistics, 33(2):17-1, 2010.

[5] Galata, V., Avesani, C., Best, C. T., Di Biase, B. and Vayra, M., “The Italian Roots in Australian Soil (IRIAS) multilingual speech corpus. Speech variation in two generations of Italo-Australians”, Language Resources and Evaluation, 56:37-78, 2022.

[6] De Iacovo, V., Mairano, P. and Hajek, J. “Does gemination resist linguistic attrition? A study on Italian migrant speech in Melbourne Australia”, in Proceedings of the 20th International Congress of Phonetic Sciences, 2860-2864, 2023.

[7] Muljačić, Ž., Fonologia generale e fonologia della lingua italiana, Il Mulino, 1969.

[8] Canepari, L., Italian pronunciation & accents: Geo-social applications of the natural phonetics & tonetics method, Lincom Europa, 2018.

[9] Romano, A., Inventari sonori delle lingue. Elementi descrittivi di sistemi e processi di variazione segmentali e sovrasegmentali, Edizioni Dell’Orso, 2008.

[10] Marotta, G. and Nocchi, N., “La liquida laterale nel livornese”, Rivista italiana di dialettologia, 25:285-326, 2001.

[11] Loporcaro, M., “Le consonanti retroflesse nei dialetti italiani meridionali: articolazione e trascrizione”, Bollettino del Centro di studi filologici e linguistici siciliani, 19:207-233, 2001.

[12] Vagges, K., Ferrero, F., Magno-Caldognetto, E. and Lavagnoli, C., “Some acoustic characteristics of Italian consonants”, Italian Journal of Linguistics, 3(1):69-84, 1978.

[13] Cox, F. and Palethorpe, S., “Australian English”, Journal of the International Phonetic Association, 37(3):341-350, 2007.

[14] Cox, F. and Fletcher, J. Australian English pronunciation and transcription. Cambridge University Press, 2017.

[15] Wells, J. Accents of English. Cambridge: Cambridge University Press, 1982.

[16] Loakes, D., Hajek, J. and Fletcher, J., “The /el/-/æl/ Sound Change in Australian English: A Preliminary Perception Experiment”, in Treis, T. and De Busser, R. (Eds) Selected Papers from the Proceedings of the 2009 Conference of the Australian Linguistic Society, 1-31, 2010.

[17] Clothier, J., “A sociophonetic analysis of/l/darkness and Lebanese Australian ethnic identity in Australian English”, in Proceedings of the 19th International Congress of Phonetic Sciences (Vol. 5), 2019.

[18] Horvath, B. M. and Horvath, R. J., “A Multilocality Study of a Sound Change in Progress: The Case of /l/ Vocalisation in New Zealand and Australian English”, Language Variation and Change, 13:37-57, 2001.

[19] Turton, D., “Sociophonetics and laterals”, in Strelluf, C. (Ed) The Routledge Handbook of Sociophonetics, 214-236. London: Routledge, 2023.

[20] Benmamoun, E., Montrul, S. and Polinsky, M. “Heritage languages and their speakers: Opportunities and challenges for linguistics”. Theoretical Linguistics, 39(3-4): 129-181, 2013. https://doi.org/10.1515/tl-2013-0009

[21] Chang, C. B. and Yao, Y., “Production of neutral tone in Mandarin by heritage, native, and second language speakers”, in Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia, 2291-2295, 2019.

[22] De Leeuw, E., Tusha, A. and Schmid, M. S., “Individual phonological attrition in Albanian–English late bilinguals”. Bilingualism: Language and Cognition, 21(2):278-295, 2018.

[23] Anwyl-Irvine, A. L., Massonnie, J., Flitton, A., Kirkham, N. and Evershed, J. K., “Gorilla in our midst: An online behavioural experiment builder”, Behaviour Research Methods, 52:388-407, 2020.

[24] Boersma, P., “Praat, a system for doing phonetics by computer”, Glot. Int., 5(9):341-345, 2001.

[25] Lennes, M. “SpeCT - Speech Corpus Toolkit for Praat (v1.0.0)”. First release on GitHub (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.375923, 2017.

[26] Traunmüller, H., “Analytical expressions for the tonotopic sensory scale”. Journal of the Acoustical Society of America, 88:97-100, 1990.

[27] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/, 2023.

Factors Influencing Flexibility in Vowel Categorisation

Stephanie Cooper, Brechtje Post

Phonetics Lab, University of Cambridge
sc2346@cam.ac.uk, bmbp2@cam.ac.uk

Abstract

Phonological category boundaries are flexible, and listeners are highly capable of interpreting variation in speech. This is particularly evident in vowel categorisation [1, 2]. Some factors that may facilitate or impede this flexibility are spectral distance from the listener's native vowel category, listener dialect, vowel frontness, and lexical bias [3, 4]. The current project explored the relative impact these factors may have on flexibility of category boundaries, to understand more about tolerance of variation. Continua of VC and CVC-structured sequences were used in a forced choice categorisation task. The stimuli varied in their lexicality; some continua contained only words or nonwords, and some contained both. The continua also varied in vowel frontness. Listeners from two different dialect groups (Standard Southern British English and New Zealand English) took part in the categorisation task. Results showed that lexical bias influences vowel categorisation, but only across an ambiguous boundary and categorising away from an underspecified vowel. This offers support for a range of theories such as Ideal Adapter [5] and Featural Underspecification [6, 7].

Index Terms: vowel categorisation, flexibility, lexical bias

1. Introduction

Listeners accommodate a range of genders, accents, and contexts in speech. They often process sounds from different speakers that may be similar in acoustic features, but in context be intended to target different phonological categories. Somehow, listeners sort variation into these discrete categories with few disruptions to communication. Phonological categories have prototypical members and boundaries, which may be "flexible" to include incoming sounds varying from the prototypes [8].

Vowel categorisation in particular appears to be more flexible than consonant categorisation. This may be because there are fewer discrete events in the spectrum of a vowel than in a consonant; it is therefore harder for listeners to hold a standard representation in their auditory short-term memory for as long [9]. Shaw and colleagues [2, 10] found that listeners were equivalent in their perception of isolated vowels in unfamiliar dialects and in their native dialect, and did not perform at ceiling in any dialect. In similar work [1], we found that New Zealand listeners successfully categorised Greek-accented English KIT[1] vowels (which are raised close to /iː/) whether or not they had prior exposure to the accent. The speech perception system is evidently capable of accepting a range of variant vowels, but it is currently unknown how far this acceptance can stretch.

Many factors may determine the limits of this flexibility. The distance in the vowel space an between an incoming vowel and the listener's native category influences how easily they will accommodate it [12]. However, in a recent experiment, we found that listeners were capable of accommodating variant vowels from a broader range of the vowel space than expected [13]. Listeners also have a strong lexical bias, in which they prefer to hear words over nonwords, and thus will preferentially categorise ambiguous sounds as phonemes leading to word comprehension [3, 4]. Listener dialect is also likely to play a role in determining category boundaries and prototypes. Other factors such as stimuli range and task structure have also been found to influence vowel categorisation [14], but they are beyond the scope of this study.

The present study set out to examine whether lexical bias, dialect spoken and vowel frontness interact with distance from native categories when listeners categorise vowels. Dialects examined were Standard Southern British English (SSBE) and New Zealand English (NZE), as these have the same phonological categories, but some differ in their placement in the vowel space [15, 16]. We also included measures from both front and back vowels, to see whether flexibility would be similar in different parts of the vowel space. We predicted that dialect groups would show peaks of categorisation where their native categories are, and that front and back vowels would show similar levels of flexibility. We also predicted that lexical bias would influence vowel selection, as no matter where on the vowel continuum, listeners would expand categories to include variant vowels when this would lead to perception of a word.

2. Method

2.1. Participants

50 monolingual participants aged 18-39 were recruited from Prolific and personal networks. They had been raised and were living in either the south of England (SE group; n = 25) or Aotearoa/New Zealand (NZ group; n = 25), without living outside of this region for more than 6 months in the past. They identified as speaking English with an accent representative of the region they grew up in and live in.

Participants were paid £5/$10NZD via Wise or Prolific upon completing the experiment.

2.2. Materials

2.2.1. Vowel continua

Eight continua consisting of a range of words and nonwords across front and back vowels were selected for use in this experiment (Table 1). These continua crossed from the highest points in the vowel chart (FLEECE and GOOSE) to lower points (TRAP and LOT). These continua varied based on lexicality; they consisted of either all words ("All"), all nonwords ("None"), words positioned medially ("Medial"), or peripherally ("Peripheral").

A male speaker of SSBE produced all points of the continua. He was raised and currently lives in the south of England.

[1] Wells' [11] lexical set vowels are used here and throughout the text, as cross-dialectal vowel perception is discussed.

Table 1: *Continua across front and back vowels used in experiment. Words are in bold. "Canonical point" indicates the points on the continuum where these canonical productions lie.*

Frontness	Lexicality	Canonical point on continuum			
		1	9	17	25
Front	All	**heed**	**hid**	**head**	**had**
	None	heeb	hib	heb	hab
	Medial	eeg	igg	**egg**	agg
	Peripheral	**feet**	**fit**	fet	**fat**
Back	All	**shoot**	**short**	**shot**	
	None	foop	forp	fop	
	Medial	oob	**orb**	obb	
	Peripheral	**hoop**	horp	**hop**	

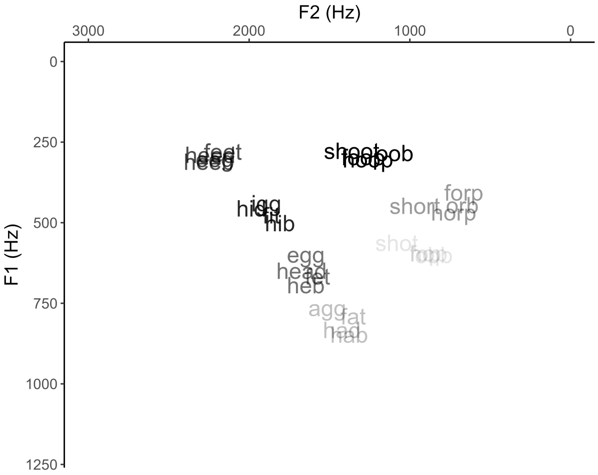


Figure 1: *F1 and F2 of words used to create continua.*

He speaks English as a native language, as well as beginner Russian and intermediate Italian.

Continua between adjacent points were created using Tandem-STRAIGHT [17] with a morphing percentage ranging from 0-100% in 9 equidistant intervals. These were combined to form larger continua between end points (e.g. between *feet* and *fat*). Due to their larger spectral range (Figure 1), front vowel continua contained 25 points between their FLEECE vowel endpoint and their TRAP vowel endpoint, while back vowel continua contained 17 points between their GOOSE vowel endpoint and their LOT vowel endpoint.

2.3. Procedure

The experiment was created using Gorilla Experiment Builder [18]. Participants were asked to use a desktop or laptop in a quiet place, with headphones to hear the auditory stimuli.

Participants read an information sheet and completed a consent form. They completed a short questionnaire to provide basic demographic information and answer questions about accents they hear in their everyday life.

The main task was forced choice, which examined how participants would classify sounds along continua between different words and nonwords. Forced choice was chosen over AXB or Go/No-Go in order to study categorisation, rather than discrimination, with multiple options available. In each trial, participants listened to a sound from a point on one of the continua, and then clicked on the screen which of the options they heard (Figure 2). While not present in the canonical productions, a PALM member of the front vowel continua was added to response options, in case participants perceived it. Response options otherwise reflected the words used to build each continuum, and varied depending on which continuum was presented.

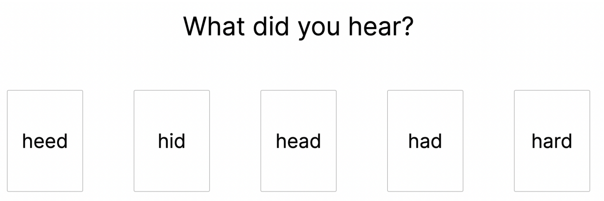

What did you hear?

Figure 2: *Example of forced choice task screen. Participants hear a sound along a continuum (heed-had in this instance), and would click one of the buttons on the screen to classify.*

The number of response options provided can affect how participants categorise sounds [14]; we chose to give participants five for front vowel continua and three for back vowel continua to show where on the continua two or more categories may be activated. We did not offer more than these options to minimise distraction from options that would likely never be perceived.

There were 168 trials in total: 25 points per continuum x 4 types of lexicality for front vowels, and 17 points per continuum x 4 types of lexicality for back vowels.

Participants also completed a perceptual acuity task to ensure that they could distinguish differences between sounds without having to classify them. Individual performances in this task and in the categorisation task were compared, and data was excluded from three participants who did not show both strong discrimination and categorisation.

2.4. Analysis

For an overview of all vowels, lexicalities were collapsed and classification curves for each vowel option were visually examined. For the purpose of this paper, only classification of DRESS and THOUGHT variants across different lexical continua were statistically analysed, as these vowels sit at interesting points between the two dialects; the SSBE DRESS vowel is near the NZE TRAP vowel, while NZE and SSBE THOUGHT vowels are more similar [15, 16]. Their placement in the centre of the continua that were used also allow the possibility of analysing the boundaries on either side of the category.

DRESS and THOUGHT data were analysed using a binomial generalised additive mixed model (GAMM) with the *mgcv* package in R [19]. GAMMs are appropriate for non-linear data and can account for both fixed and random effects [20].

The model was formed in a stepdown manner, in which a full model with all possible parametric and smooth terms was built, and then compared to models with one term removed at a time. AIC scores and the compareML function from the *itsadug* package in R [21] were used to compare models.

3. Results

3.1. All vowels

Figure 3 shows categorisation of points on the continua as different vowels (collapsed across all levels of lexicality) represented by members of Wells' lexical set.

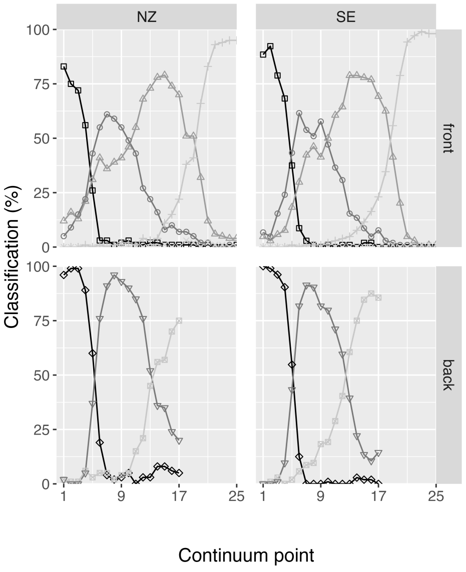


Figure 3: *Categorisation of points along front and back vowel continua as different vowel variants by NZ and SE listeners. Variants are represented by lexical sets. X axis marks indicate canonical productions by male SSBE speaker.*

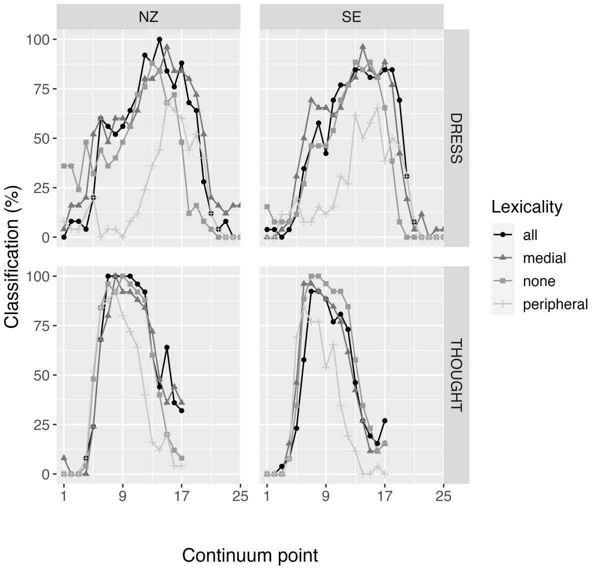


Figure 4: *Acceptance of points on the respective front or back vowel continua as containing* DRESS *or* THOUGHT *vowels by NZ and SE listeners. Responses to different continua with different levels of lexicality are indicated.*

The Dialect groups do not appear to differ greatly in their position and sizes of the categories. In both groups, the vowel categories are not very distinct and boundaries are blurred, both for front and back vowels. The KIT category in particular has overlap with the FLEECE and DRESS categories, and even at its peak of categorisation it is only chosen around 60% of the time.

The steeper the gradient of the curve, the more distinct categories can be assumed to be. The boundary between GOOSE and THOUGHT, for example, appears to be more distinct than between THOUGHT and LOT, as the shift in classification between the vowels occurs over about 4 steps from GOOSE to THOUGHT compared to about 7 steps from THOUGHT to LOT.

3.2. DRESS vowels by lexicality

Figure 4 shows the classification of points as DRESS vowels, with the lexicalities of the vowel continua represented as different lines. A GAMM was fitted to represent this data, and the best fit contained an interaction between dialect spoken and continuum lexicality as a parametric term and as a smooth term, and participant included as a random smooth term.

Visual examination of the model using difference smooth plots for different Dialect*Lexicality conditions shows that for both Dialect groups (example in Figure 5), the Peripheral condition has the most distinctly different shape compared to other lexicality conditions. It differs significantly ($p < 0.05$) from All, None and Medial vowel continua conditions between points 5 and 20, outside of which classification as a DRESS vowel decreases for all lexicality conditions. The Peripheral condition produces a shallower curve with a lower peak of DRESS classification. The Dialect groups are similar to each other in the shape and position of their other lexicality conditions as well.

3.3. THOUGHT vowels by lexicality

Figure 4 also shows the classification of points along the vowel height spectrum as THOUGHT vowels by lexicality of continua. A GAMM was fitted to represent this data, and again the best fit contained an interaction between dialect spoken and continuum lexicality as a parametric term and a smooth term, and with participant included as a random smooth term.

Again, visual examination using difference smooth plots for Dialect*Lexicality conditions showed that the Peripheral condition curve is the most distinctly different for both Dialect groups. It differs significantly ($p < 0.05$ in all instances) from All, None and Medial vowel continua conditions between points 6 and 17, with the exception of the NZ graph where it differs significantly from the None condition only until point 16. Similarly to the DRESS curves, the difference between the Peripheral conditions and others is asymmetrical. In the THOUGHT curves, the Peripheral condition shows a steep increase in classification along the GOOSE-THOUGHT boundary, as the other conditions

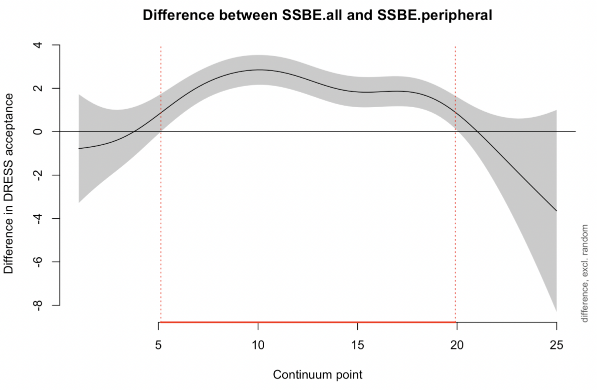


Figure 5: *Example difference smooth plot, here showing difference between SSBE.all and SSBE.peripheral smooths from fitted front vowel model.*

do. However, THOUGHT classification then drops off earlier along the THOUGHT-LOT boundary for the Peripheral condition than it does for the other lexicalities.

The NZ group's back vowel Medial condition does seem to show some difference from the other Medial conditions, as classification of the variants as "orb" stays high throughout the second half of the continuum. However, this also applies to the back vowel All condition, where classification as "short" also stays high. Apart from this NZ-specific pattern, the Dialect groups are again similar in their classification shapes and positioning along the continua.

4. Discussion

This study was an initial investigation into some of the factors that influence flexible vowel categorisation. Participants heard sounds across a range of vowel continua and classified what they heard as a word or nonword option. Similar to previous findings of vowel flexibility [1, 2, 13], we found that at least in English, with its large vowel inventory and wide range of dialects, vowel perceptual categories are not clearly defined and bleed into each other. One given vowel sound may be classified as a member of multiple categories depending on the context it is heard in. However, some category boundaries appear to be more distinct than others.

In the Peripheral lexicality condition, listeners appear to classify points between the two categories as the one which will elicit perception of a word over a nonword, but only across indistinct boundaries. This occurs despite the mid vowels in the Peripheral condition having very similar acoustic qualities to those in the other conditions (Figure 1), thus reducing the likelihood that this can be attributed to other factors like the broad range of the stimuli [14]. Any shifts in classification can therefore be attributed to lexical effects. These top-down influences are most effective when listeners are uncertain about categorisation, as they are more evident across indistinct category boundaries. These influences are also transient and do not generalise to classification across word-word or nonword-nonword boundaries, as similar results for the same categories were not found in All or None lexicality conditions. Lexicality effects were expected, as they are quite pervasive [4], but the lack of effect in the Medial condition suggests an interaction between lexicality and blurry boundaries, as addressed further below.

It is interesting to note that despite documented differences in the vowel spaces of SSBE and NZE dialects, we found minimal difference between groups in the position of their vowel classifications. Additionally, for both groups, the peaks of KIT and DRESS vowels were at a slightly more raised point on the vowel continua (shifted to the left) than the canonical productions by the SSBE speaker. This was expected for the NZ group, who tend to have more raised front vowels [16], but is surprising for the SE group, who would be expected to classify vowels in a way consistent with the SSBE productions. We are currently replicating this experiment with a NZE speaker to further examine possible effects of speaker and listener dialect on these results. One explanation may be that the speaker's vowels are slightly influenced by other languages he speaks [22].

The NZ listeners reflecting SE classification patterns may also be due to the use of a SSBE speaker in stimuli creation. When asked in a concluding questionnaire what accent they thought the speaker had, NZ participants overwhelmingly responded with "British" or "English". Even though the continua extended across the vowel space, so any point could have been a canonical production from any accent, segments like the non-centralised KIT vowel and the low TRAP vowel may have led NZ listeners to correctly identify the accent as different to theirs and like SSBE, which they have a relatively high level of familiarity with. This is consistent with frameworks such as Ideal Adapter [5] which suggests that where possible, listeners select a model based on previous experience to help them to decode the incoming speech signal. NZ listeners in this experiment may have used the salient features above to select an SSBE model to aid in the task, thus shifting their vowel categorisation to be consistent with this model. Our replication with an NZE speaker will also reveal further information about this theory.

These results support previous findings that mid vowel categories are particularly capable of accommodating variation. In a comparison of NZ and American listeners, NZ listeners accepted more variation in their TRAP vowel than American listeners [6]. This was proposed to be due to TRAP being raised to a mid vowel in NZE, and thus lacking the feature specification of [+high] or [+low]. The authors advocate for a Featurally Underspecified Lexicon [7] in which some segments are defined by the absence of features, and thus may be more accepting of variation. The current study did find that FLEECE and TRAP seem to have more distinct vowel boundaries than DRESS, as indicated by their steeper classification curves. NZ listeners did not show as much range in their acceptance of TRAP vowels as found previously, which may again be due to participants identifying the speaker as British and shifting TRAP classification lower.

Featural underspecification may also provide an explanation for lexical effects in the Peripheral but not Medial condition. DRESS and THOUGHT categories are in the middle of the vowel space, so are likely to be the least specified for height. While in other circumstances this may lead to higher acceptance of variation, in the Peripheral condition when classification as a mid vowel led to nonword perception, the underspecification may have shifted classification in the direction of the adjacent category leading to perception of a word (i.e. KIT or LOT). The inverse may not be true; in the Medial condition where classification as KIT or LOT led to a nonword, their height specification may have been strong enough to resist a shift towards DRESS or THOUGHT, despite the lexical benefit of choosing a mid vowel in this condition. The exception to this is in the NZ Medial back vowel continuum, where some THOUGHT classification seems to extend further towards LOT than it does in other conditions (although the All condition shows a similar pattern). This may indicate underspecification of the LOT category in NZE, which is more of a mid vowel than it is in SSBE. It appears that lexical bias and category specifications interact to influence ambiguous vowel classification.

There is much further analysis of this data to be completed. We plan to look more at how lexicality affects other vowels with substantial overlap, like KIT, but also more specified vowels like FLEECE or GOOSE. We also hope to work with EEG to examine differences in neural responses when variants are accepted or rejected as members of a phonological category under different lexical conditions. Previous research has used differences in the N400 or other neural responses to examine how and when variation is accommodated in speech processing [23, 24].

The aim of this study was to explore the weighting of some factors that influence vowel categorisation flexibility. We found that across front and back vowels for two dialect groups, lexicality and specificity of vowel categories influence categorisation curves along a height continuum. These factors also interact in that underspecified categories are more susceptible to lexical bias. This provides a useful foundation for examining vowel flexibility further, and offers many avenues for future research.

5. References

[1] Cooper, S. and Cooper, S., "Exposure-independent comprehension of Greek-accented speech: evidence from New Zealand listeners," in Proceedings of the 20th International Congress of Phonetic Sciences, 6–10, 2023,

[2] Shaw, J. et al., "Resilience of English vowel perception across regional accent variation," Laboratory Phonology, 9(1), 1–36, 2018.

[3] Fox, R.A., "Effect of lexical status on phonetic categorization," Journal of Experimental Psychology: Human Perception and Performance, 10(4), 526–540, 1984.

[4] Gaskell, M.G. and Marslen-Wilson, W.D., " Mechanisms of phonological inference in speech perception", Journal of Experimental Psychology: Human Perception and Performance, 24(2), 380–396, 1998.

[5] Kleinschmidt, D.F. and Jaeger, T.F., "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel," Psychological Review, 122(2), 148–203, 2015.

[6] Scharinger, M. and Idsardi, W. J., "Sparseness of vowel category structure: Evidence from English dialect comparison," Lingua, 140, 35–51, 2014.

[7] Lahiri, A. and Reetz, H., "Distinctive features: Phonological underspecification in representation and processing," Journal of Phonetics, 38(1), 44–59, 2010.

[8] Xie, X., Theodore, R.M. and Myers, E.B. "More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories," 43(1), 206–217, 2017.

[9] Pisoni, D.B., "Auditory and phonetic memory codes in the discrimination of consonants and vowels," Perception & Psychophysics, 13(2), 253–260, 1973.

[10] Shaw, J.A. et al., "Revealing perceptual structure through input variation: cross-accent categorization of vowels in five accents of English," Laboratory Phonology, 14(1), 2023.

[11] Wells, J.C., Accents of English: Volume 1. Cambridge University Press, 1982.

[12] Best, C.T., "The emergence of native-language phonological influences in infants: A perceptual assimilation model," in The development of speech perception: The transition from speech sounds to spoken words, 167–224, 1994.

[13] Cooper, S. and Post, B. "Flexibility of vowel categorisation in newly learned words", Colloquium of British Association of Academic Phoneticians, Cardiff, 2024.

[14] Benders, T., Escudero, P. and Sjerps, M.J. "The interrelation between acoustic context effects and available response categories in speech sound categorization", The Journal of the Acoustical Society of America, 131(4), 3079–3087, 2012.

[15] Hughes, A., Trudgill, P. and Watt, D. English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles, Fifth Edition, Routledge, 2012.

[16] Bauer, L., Warren, P., Bardsley, D., Kennedy, M. and Major, G. "New Zealand English", Journal of the International Phonetic Association, 37(1), 97–102, 2007.

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

[18] Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N. and Evershed, J.K. "Gorilla in our midst: An online behavioral experiment builder," Behavioural Research, 52(1), 388–407, 2020.

[19] Wood, S. "mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation", 2023, available at https://cran.r-project.org/web/packages/mgcv/index.html.

[20] Sóskuthy, M., "Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction," arXiv, 2017, http://arxiv.org/abs/1703.05339.

[21] van Rij, J., Wieling, M., Baayen, R.H. and van Rijn, H. "itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.", 2022. https://cran.r-project.org/web/packages/itsadug/index.html

[22] Flege, J. "Second language speech learning: Theory, findings and problems," in Speech Perception and Linguistic Experience: Issues in Cross-Language Research, 229—273, Strange, W. (Ed.), York Press, 1995.

[23] Sebastián-Gallés, N., Vera-Constán, F., Larsson, J.P., Costa, A. and Deco, G. "Lexical plasticity in early bilinguals does not alter phoneme categories: II. Experimental evidence," Journal of Cognitive Neuroscience, 21(12), 2343–2357, 2009.

[24] Politzer-Ahles, S., Lin, J., Pan, L. and Lee, K.K. "N400 Evidence That the Early Stages of Lexical Access Ignore Knowledge About Phonological Alternations", Language and Speech, 65(2), 354–376, 2022.

Visualising Vowel Dynamics for the Training of Text to Speech Systems

Henry An, *Jesin James*, *Catherine Watson*, *Binu Abeysinghe*

Department of Electrical, Computer and Software Engineering, University of Auckland

`han285@aucklanduni.ac.nz, jesin.james@auckland.ac.nz, c.watson@auckland.ac.nz, babe269@aucklanduni.ac.nz`

Abstract

Currently, the main tool used for visualising text to speech training are learning curves. Recent research examined the use of linguistic features for the visualisation of the training of speech synthesis systems by plotting the monophthongs of the model by F1 and F2 measurements at the midpoint and confirmed its applicability. This paper investigates a proposed approach for improving linguistic based visualisation for the training of text to speech systems by representing formant dynamics in synthetic speech vowels. A model was fine tuned from American English to New Zealand English and this process was used to test two visualisation designs.

Index Terms: speech synthesis, machine learning, formant trajectory

1. Introduction

The field of text to speech has advanced to a stage where there exists synthetic speech such that the difference in naturalness between synthetic and natural speech is statistically insignificant [1]. This advancement is in part driven by the adoption of neural network based synthetic speech approaches. Neural network based text to speech (TTS) are a type of machine learning based speech synthesis where a deep neural network model is trained on a dataset and then can be fed input text to produce speech [2].

As the technology of text to speech matures, its applications have become more widespread and, for example are used in assistant systems like Siri or Alexa [3] but also for purposes such as screen readers or to give pronunciation examples for language education [4]. As such, for the purposes of equity, it is important that text to speech systems are available in a variety of languages and dialects and thus important to be able to effectively train TTS models for a variety of languages and dialects.

In machine learning, a major form of training visualisation used are learning curves. Learning curves plot the progress of a metric over the course of the training process. A typical metric would be the loss function, which represents the error between the predictions made by the model compared to the actual data provided. These learning curves can be used to visualise the progress of the training, representing, for example, how well the model is learning the data. However, in the case of TTS, this is not fully sufficient. In TTS training we not only need to see how well the model is learning the information of a dataset through the visualisation of metrics such as loss, but also what the synthesised speech of the model sounds like to humans. To this end using a linguistic based visualisation can be helpful [5].

Vowels are one of the types of sound which are found in languages. They are the sounds which are formed by an unimpeded flow of air through the vocal tract and have identifying features such as formants. The formants of a vowel are the resonances of the vowel at certain frequencies, of which formant 1 (F1) and formant 2 (F2) are considered the most important for identifying the vowel [6]. The vowels of a language can be said to form a vowel space which is the space occupied by the set of vowels of said language in terms of a two dimensional F1, F2 space [7]. By having access to the vowel space, one can visually identify the pronunciation of vowels in a sample of speech. A common way of plotting a vowel space is by taking a measurement of F1 and F2 at the midpoint of the vowel duration. These points can then be plotted on a two dimensional frequency graph.

In the study by Abeysinghe et al, a model was trained from American English to New Zealand English. It was observed that the vowel space shifted from one resembling the original American English to one resembling the desired New Zealand accent [5]. If samples of speech from the American and New Zealand English models are synthesised and then the vowel spaces are plotted, one will be able to see the differences between the two models visually by comparing the vowel spaces and thus see the training progress of the model in terms of pronunciation. This study focused on monophthongs and thus relied on measurements of the vowel formants taken at the midpoint only. However, this does not form a complete picture of the vowel pronunciation as formant values can change over the duration of a vowel, even for monophthongs, and can be important for identifying the vowel [8].

This research examines how to improve this visualisation, mainly by extending the visualisation to contain dynamic aspects of the vowel. Previous work [5] established the validity of a linguistics based approach in the visualisation of TTS model training by examining the effectiveness of plotting the F1 and F2 at the midpoints of the vowels. By only taking formant measurements at the midpoint, only a snapshot of the vowel is captured and used to represent the vowel. This can be termed the *static visualisation of the vowels*. But the formants of the vowel do not necessarily stay the same for the whole duration of the vowel and this change, termed *formant dynamics*, has been used to analyse the historic sound change of vowels, including monophthongs [9] [10]. In some cases, such as for diphthongs, this movement is an identifying feature of the vowel. Thus, in order to more fully visualise the vowels of a model, in particular said diphthongs, this paper explores two ways of plotting formant dynamics for a TTS training context. To guide the investigation into improving this visualisation, we developed the following research questions:

RQ1: How best can formant dynamics be represented for the training of TTS systems?

RQ2: How useful is formant dynamic information to the person training TTS models?

In the rest of this paper, we will explain our methodology and discuss the results. Finally, a conclusion will be drawn, and future work discussed.

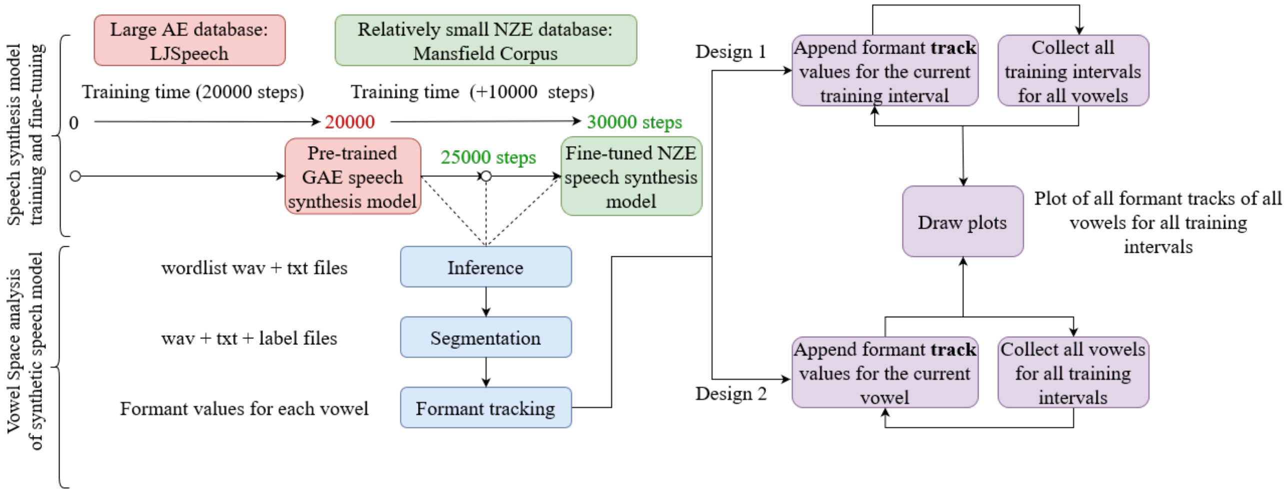

Speech synthesis model training and fine-tuning
Large AE database: LJSpeech
Relatively small NZE database: Mansfield Corpus
Training time (20000 steps)
Training time (+10000 steps)
0
20000
30000 steps
Pre-trained GAE speech synthesis model
25000 steps
Fine-tuned NZE speech synthesis model
Vowel Space analysis of synthetic speech model
wordlist wav + txt files
Inference
wav + txt + label files
Segmentation
Formant values for each vowel
Formant tracking
Design 1
Append formant track values for the current training interval
Collect all training intervals for all vowels
Draw plots
Plot of all formant tracks of all vowels for all training intervals
Design 2
Append formant track values for the current vowel
Collect all vowels for all training intervals

Figure 1: *Flow chart of the visualisation implementation across training. Adapted from [5] .*

Table 1. *Table of diphthongs of New Zealand English*

Lexical Set [11]	FACE	PRICE	CHOICE	MOUTH	NEAR	GOAT
Synthesised Word Used	aid	hide	lloyd	how’d	hare	oat
Truth Word Used	date	hide	voice	about	air	code
IPA Transcription	æɪ	ɑe	oɪ	æɔ	eː	əʉ

2. Methodology and Implementation

This research, based on the previous static formant visualisation, focused on modifying the system to include dynamic formant information and then investigating the merits of different visualisations. To answer our research questions, we first trained a model on American English for 20k steps before further training on New Zealand English for 10k total steps for a total of 30k. We then analysed the formants of the vowels of the models and produced two designs for visualising the formants. This process of training the model and drawing plots at select stages can be seen in Figure 1.

In linguistics, one way formant dynamics are represented are formant trajectories [12]. Formant trajectories are made by taking measurements of the F1 and F2 values of a vowel at multiple fixed points in the vowel duration before plotting the trajectories in a F1, F2 frequency graph. The designs of the systems were inspired by the usage of formant trajectory plots which are used to visualise formant dynamics in the field of linguistics such as in [12].

While the previous approach visualised all of the vowel spaces of all of the training stages in a single plot, this was not the method taken in our approach. This was because our study investigated visualising the vowel dynamics instead of the vowel spaces and as such there was more information to be represented. Thus, two designs were created which presented the vowel trajectories in different ways and compared.

The two designs were investigated to understand how best to visualise the dynamic formant information and were compared, hereafter referred to as design 1 seen in Figure 2 and design 2 seen in Figure 3. Design 1 attempted to group the vowels by phoneme and plotted each vowel phoneme as an individual plot with each curve representing the vowel of an individual model while design 2 was to plot the vowel trajectories of each training stage and juxtapose the plots next to each other, with each curve representing an individual diphthong in the vowel space of a model.

The investigation began by training a model to experiment on. The FastSpeech2 [13] system was used for this purpose. In order to train a model on FastSpeech2, as seen in figure 1, text grids with the phonetic alignments need to be generated of the speech data. This was done automatically using the Montreal forced aligner system [14]. The data used to train the model consisted of two sets. One was the LJSpeech corpus [15], a dataset of a female speaker of American English, which was used to train the base model and the second was a sample of a female speaker of New Zealand English from the Mansfield Corpus [16] which was used to fine tune.

The implementation of FastSpeech2 used was based on an implementation made using PyTorch [4] which itself was based on [17]. The proposed idea was to interrupt the training process at fixed intervals of steps, such as when a model was saved, and create a visualisation of the vowel space of the TTS model at each saved step with the formant trajectory of the vowels included. Samples of speech from the model were generated using words of a ’h’ vowel ’d’ format (hVd words) as the input when possible and fed into a WEBMAUS [18] forced aligner API to generate phonetic alignment text grids. Then the generated alignments were processed to extract the vowels, for which each vowel, for each training stage had the formant data from the beginning, middle, and end extracted. Then the data is assembled into a Numpy data frame [19] which is used to plot the graph.

The words used to generate the formant plots as well the ground truth, the words of the New Zealand English speech used to train the model, are listed in Table 1 in addition to their IPA transcriptions. Not all the words are hVd words, as in the case of the synthesised speech, hVd words could not be reliably generated, and in the case of the ground truth, was not available.

Table of Contents for this manuscript

Figure 2: *Visualisation design type 1, where each curve represents the trajectory of a vowel in an individual model. (Splines are visualisation aids, only the 3 indicated points are real measurements)*

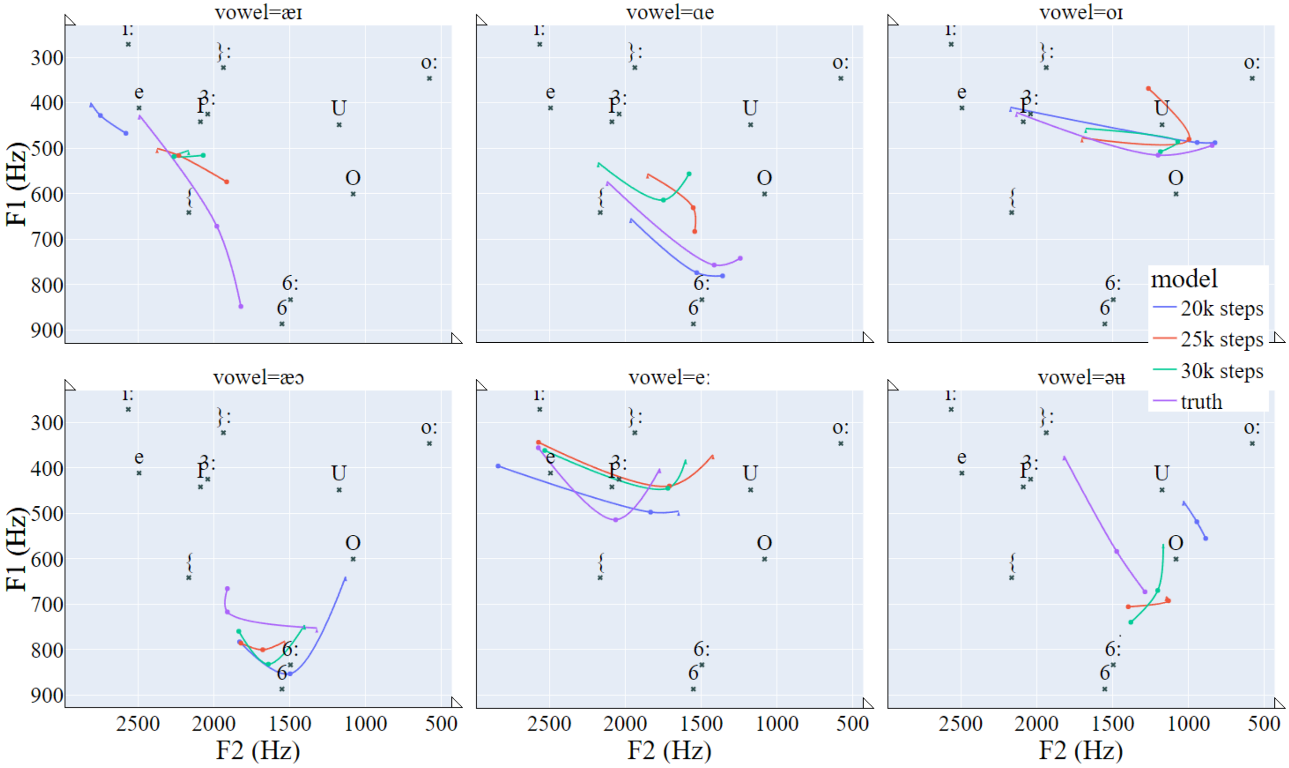
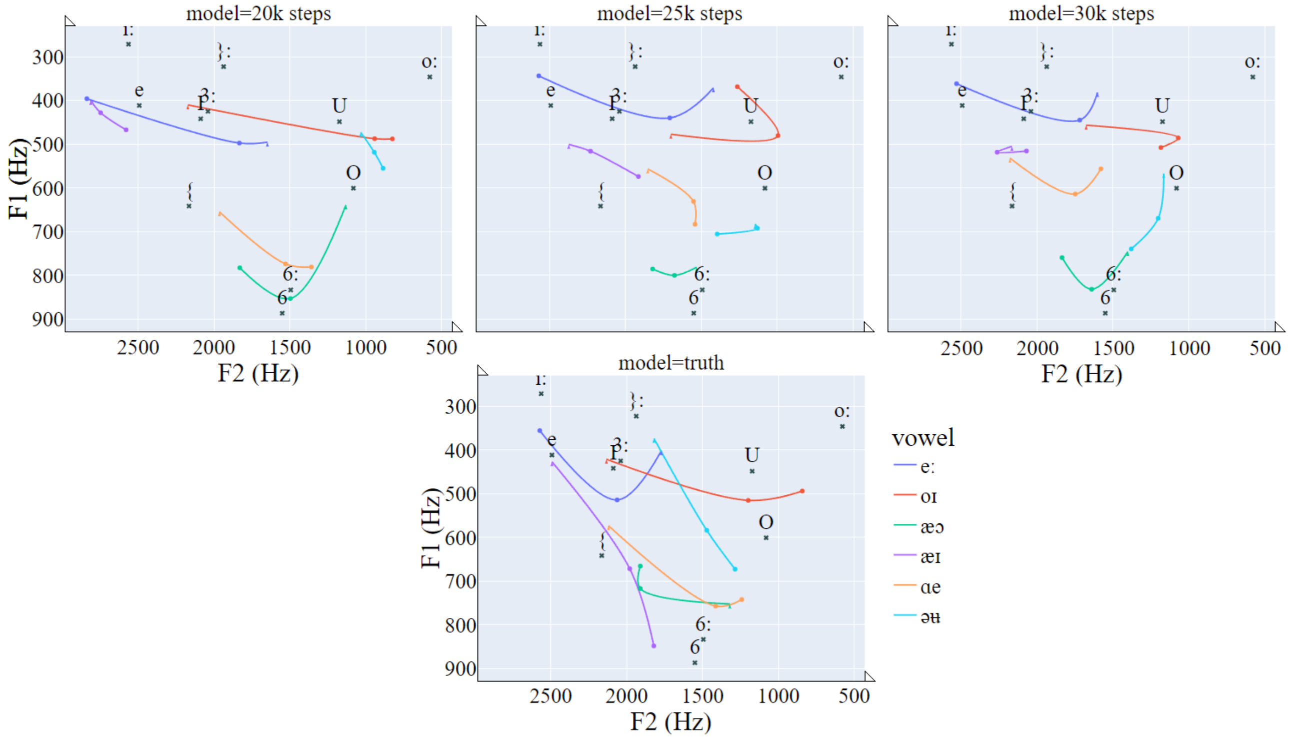


Figure 3: *Visualisation design type 2, where each curve represents the trajectory of an individual vowel in the vowel space of a particular model. (Splines are visualisation aids, only the 3 indicated points are real measurements)*



The lexical set of the word from [11] is also given.

The two designs plot the formant trajectory of each vowel in a F1, F2 space. The formants are plotted by frequency and the formant trajectories are created using frequency measurements from the start, middle and end of the vowel duration. Measurements for the start and end are taken at 20 percent from the end and beginning to mitigate the coarticulatory effects at the edges of the vowel while the measurement of the middle is taken at the exact midpoint of the vowel duration. The system was created with Python using the Parselmouth [20] library which is an implementation of Praat [21] functions within Python. Parselmouth is used to extract the formant data which is then plotted using Plotly [22].

3. Results and Discussion

Here in Figures 2 and 3 we can see a model being fine tuned from 20k steps trained American base model for 10k steps to a 30k steps total trained New Zealand English model. We can see that both designs consist of a series of F1, F2 plots of the formant trajectories of the vowels analysed for all of the training stages analysed. The two designs are differentiated by how they group the vowel trajectories in different ways. Design 1 groups the trajectories by the vowel phoneme while design 2 groups the trajectories by training stages of the model. The differences between the set of vowel trajectories of each training stage is emphasised on the training stage plots while the progression of the vowels over the training process is emphasised on the vowel phoneme graphs. In both designs, the vowel is represented by 3 points which have been connected with a line where the end representing the beginning of the vowel is unmarked and the side representing the end of the vowel is marked with an arrowhead. The lines were smoothed using a spline to give an approximation of the formant trajectory. The spline serves only an illustrative purpose and is not a prediction of the formant values.

By plotting the formants trajectories of each vowel rather than representing vowels as F1, F2 frequencies taken at their midpoint, the change in F1 and F2 over the duration of the vowel is also visualised. This can help improve the visualisation of the changes to vowel production that occur when training a TTS. An example can be seen in the vowel æɪ in 3. The midpoints of the 25k steps and 30k steps models occur near each other which would make them appear very similar in a single point representation, with a distance of only 32Hz, but when the start and end points are added, the difference between the two models becomes more clear with the star points being 165Hz apart and the end points being 208Hz apart. The greater the distance, the greater the difference in vowel quality, in effect indicating how similar two vowels sound. When comparing to the ’truth’ this would indicate the closeness of the model to the desired voice.

As formant dynamics are useful for the analysis of the historic change in vowels [10], it stands to reason that this can also be applied to analysis of change in vowels for TTS training. The two designs can also be considered to constitute two different approaches to visualisation, with design 1 being an more engineering focused perspective, examining the change in vowels, while design 2 could be considered a more linguistic based perspective, examining the shape of the vowel space as a whole. A possible usage would be to examine design 2 for a holistic understanding of the change in vowel pronunciation and to refer to design 1 when desiring further information on specific vowels.

A limitation of this approach is that it relies on automated systems to generate the text grids which cannot be hand corrected. Studies [23] have shown that while forced aligners can be highly accurate and reliable, for non American speech they still generally perform at a level lower than humans, especially for highly divergent local varieties. Thus, the visualisation cannot be seen as a guaranteed completely accurate representation of the vowel space and it would be more appropriate be used to direct the researcher to the relevant locations for manual inspection.

4. Conclusion and Future Work

As TTS has become more widespread and important, so too does the training of TTS systems. Currently, the usage of linguistic based visualisation of the training of TTS systems is in its early stages with its viability being demonstrated with static formant measurements. This paper examined two proposed approaches to extend this visualisation to formant dynamics and suggests some possible uses for such a system.

While work has been done into the investigation of formant dynamic based visualisation for the training of text to speech systems, further testing is required to draw concrete conclusions. An evaluation of the system has yet to be conducted and has been planned to consist of a survey directed at the intended users of the system. Testing of the extent to which formant dynamics are important for synthetic speech could also be conducted.

In terms of future work for TTS training visualisation, further investigation of linguistic feature visualisation could be conducted such as visualisation of higher formants, fundamental frequency or prosody.

5. References

[1] Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., and others,, “Naturalspeech: End-to-end text-to-speech synthesis with human-level quality,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] Zen, H., Senior, A., and Schuster, M., “Statistical parametric speech synthesis using deep neural networks,” in *2013 ieee international conference on acoustics, speech and signal processing*, pp. 7962–7966, IEEE, 2013.

[3] Hoy, M. B., “Alexa, siri, cortana, and more: an introduction to voice assistants,” *Medical reference services quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[4] Pine, A., Wells, D., Brinklow, N., Littell, P., and Richmond, K., “Requirements and motivations of low-resource speech synthesis for language revitalization,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7346–7359, Association for Computational Linguistics, 2022.

[5] Abeysinghe, B. N., James, J., Watson, C., and Marattukalam, F., “Visualising model training via vowel space for text-to-speech systems,” in *Proc. Interspeech 2022*, pp. 511–515, 2022.

[6] Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J., “An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns,” *Word*, vol. 8, no. 3, pp. 195–210, 1952.

[7] Klein, W., Plomp, R., Pols, L. C., and others,, “Vowel spectra, vowel spaces, and vowel identification,” *Journal of the Acoustical Society of America*, vol. 48, no. 4, pp. 999–1009, 1970.

[8] Watson, C. I. and Harrington, J., “Acoustic evidence for dynamic formant trajectories in australian english vowels,” *The Journal of the acoustical society of America*, vol. 106, no. 1, pp. 458–468, 1999.

[9] Winn, M. B. and Wright, R. A., “Reconsidering commonly used stimuli in speech perception experiments,” *The Journal of the Acoustical Society of America*, vol. 152, no. 3, pp. 1394–1403, 2022.

[10] Cox, F., Penney, J., and Palethorpe, S., “Australian english monophthong change across 50 years: Static versus dynamic measures,” *Languages*, vol. 9, no. 3, p. 99, 2024.

[11] Wells, J. C., *Accents of English: Volume 1*, vol. 1. Cambridge University Press, 1982.

[12] Renwick, M. E. and Stanley, J. A., “Modeling dynamic trajectories of front vowels in the american south,” *The Journal of the Acoustical Society of America*, vol. 147, no. 1, pp. 579–595, 2020.

[13] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y., “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *arXiv preprint arXiv:2006.04558*, 2020.

[14] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M., “Montreal forced aligner: Trainable text-speech alignment using kaldi.,” in *Interspeech*, vol. 2017, pp. 498–502, 2017.

[15] Ito, K. and Johnson, L., “The lj speech dataset.” `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[16] Watson, C. I. and Marchi, A., “Resources created for building new zealand english voices,” in *Proc. 15th Australas. Int. Conf. Speech Science and Technology*, pp. 92–95, 2014.

[17] Chien, C.-M., Lin, J.-H., Huang, C.-y., Hsu, P.-c., and Lee, H.-y., “Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8588–8592, 2021.

[18] Kisler, T., Reichel, U., and Schiel, F., “Multilingual processing of speech via web services,” *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[19] Oliphant, T. E. and others,, *Guide to numpy*, vol. 1. Trelgol Publishing USA, 2006.

[20] Jadoul, Y., Thompson, B., and de Boer, B., “Introducing Parselmouth: A Python interface to Praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[21] Boersma, P. and Van Heuven, V., “Speak and unspeak with praat,” *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[22] Inc., P. T., “Collaborative data science,” 2015.

[23] MacKenzie, L. and Turton, D., “Assessing the accuracy of existing forced alignment software on varieties of british english,” *Linguistics Vanguard*, vol. 6, no. s1, p. 20180061, 2020.

Positional Allophony, Ethnolectal Variation and /l/ Darkness in Australian English

Conor Clements, Joshua Penney, Felicity Cox

Centre for Language Sciences, Department of Linguistics, Macquarie University

conor.clements@hdr.mq.edu.au, joshua.penney@mq.edu.au, felicity.cox@mq.edu.au

Abstract

English laterals exhibit acoustic variability, often described as having "clear" or "dark" allophones determined by syllable position. Previous accounts suggest that Australian English /l/ is dark in all positional contexts, but this is understudied, and it is unclear how speaker background may affect lateral production. This study presents an acoustic analysis of /l/ produced in citation-form in different syllable positions by speakers with various heritage language backgrounds. Results show that speakers from all language backgrounds tested produce a positional distinction between clear and dark /l/, but that those with English-only backgrounds produced darker syllable-initial /l/s compared to those with non-English speaking backgrounds.

Index Terms: lateral darkness, sociophonetics, acoustics, ethnolectal variability, Australian English, allophony.

1. Introduction

1.1. Positional allophony of /l/ in English

Variation in the formant structure of the English alveolar lateral approximant is documented in many varieties [1, 2]. Broadly speaking, this variation is a continuum ranging from "clear" to "dark", with clear variants exhibiting higher F2 and lower F1 than dark variants. This variability is related to the extent of velarisation/pharyngealisation as well as the relative timing of apical and dorsal gestures involved in /l/ articulation, with clear /l/ characterised as "/i/-like" and dark /l/ as "/u/-like" respectively [1, 3, 4].

The distribution of clear and dark variants is usually understood to vary by syllable position, with clear /l/ occurring syllable-initially and dark /l/ syllable-finally, but as [5] shows the patterning of positional allophony differs between English varieties: some have notably clear onset /l/s contrasting with dark coda /l/s; others have intermediately dark onset /l/s contrasting with yet darker coda /l/s; others still entirely lack positional distinctions [6, 7, 8]. The picture is further complicated by sociolinguistic factors such as speakers' heritage language background, gender, and social class, which may additionally affect /l/ realisation [6, 9, 10, 11].

Laterals are involved in various coarticulatory processes in English that have attracted considerable attention in the literature, with studies finding both that darker /l/s are more resistant to coarticulatory influence from other vowels than lighter /l/s, and that darker /l/s may induce coarticulatory changes in adjacent vowel segments [3, 12, 13]. In Australian English (AusE), impressionistically described as having dark /l/s in all contexts [14, 15], postvocalic dark /l/ is implicated in various spectral contrast reductions, including the vowels in pairs like *feel-fill*, *fool-full*, *dole-doll*, *howl-Hal* and (for some Victorian speakers) *celery-salary* [16, 17]. Overall, the influence of coarticulatory processes in progressing sound change is well-documented [e.g., 18], but /l/ darkness as a potential contributor to this is understudied in AusE.

Though limited, some acoustic research exists on positional variability of /l/ in AusE. [19] finds that Victorian AusE speakers produce significantly darker coda /l/s than onset /l/s. Looking at intervocalic /l/s, [20] found that only 1 of their 13 participants consistently produced clear /l/s in this context. Research on /l/ darkness in AusE using data from informants who reside outside of Victoria is scarce. Additionally, we lack understanding of the acoustic characteristics of /l/ in a broad variety of positional contexts in AusE. We therefore seek to address these gaps in the current analysis.

1.2. Ethnolectal variability of /l/ darkness

Variation in /l/ darkness associated with speaker heritage language background is well-described in literature on several English dialects outside Australia. Some of this work is concerned with the acquisition of canonical English positional allophony among bilingual children [21, 22, 23], whereas other research focuses on understanding differences in positional allophony as an example of ethnolectal variation [9, 10, 11, 19, 24]. For example, literature from the United Kingdom has consistently found that speakers of South Asian heritage (e.g. with heritage languages such as Punjabi and Sylheti) tend to exhibit a canonically English-like pattern of /l/ positional allophony but with "hyper-clear" /l/ in onset contexts, even in areas where the local variety has dark /l/ onset /l/s, thereby producing a much more marked distinction between onset and coda /l/ variants [10, 11, 24]. It is presumed that this stems from heritage language transfer, but its documented persistence through multiple generations of speakers, including those who do not report any use of a heritage language, may suggest that a supralocal British Asian type of /l/ positional allophony has developed that coexists with local patterns used by Anglo-heritage speakers [9, 23, 24].

In the Australian context, [19] investigated /l/ darkness in adult AusE speakers of both Anglo-Celtic heritage and Lebanese heritage, finding that both groups produced a positional distinction between onset and coda /l/, as expected. This study made use of a social network score that measured the extent to which a Lebanese-Australian informant's social network comprised of Lebanese versus non-Lebanese members, revealing additional insights: male speakers with more densely Lebanese social networks produced clearer onset /l/s, and female speakers with more densely Lebanese social networks produced both clearer onset /l/s and darker coda /l/s. This suggests that /l/ darkness carries broader socioindexical associations reflective of a speaker's attitudes towards their heritage language background that go beyond simple effects of language transfer, and which are modulated by gender [19, 25].

Ethnolectal variation in /l/ darkness has otherwise received little inquiry in the AusE literature, presenting an interesting gap to investigate given the comparatively small body of work on this topic compared to that in other Anglophone countries. The linguistic makeup of Australian communities has changed substantially in recent decades, with cities such as Sydney having become particularly diverse [26]. Recent census data show that 64% of Greater Sydney residents have at least one parent born overseas, with languages other than English used in 42% of households [27]. Investigating /l/ darkness in the context of this increase in cultural and linguistic diversity is also key in motivating the present analysis.

1.3. Research questions and predictions

This study seeks to examine the following:

1) How do the acoustic qualities of /l/ vary in different positional contexts in AusE?
2) Is there evidence that AusE /l/ is dark in all positional contexts, as has been previously claimed?
3) Do patterns of positional allophony vary between speakers with different heritage language backgrounds?

We seek to answer these questions through an acoustic analysis of lateral variation in AusE. We first expect that syllable-initial /l/ will be clearer than syllable-final /l/ [19]. It is more difficult to predict how word-medial laterals will behave since there is less acoustic evidence available for these contexts, though some research suggests they behave more similarly to onset /l/s than coda /l/s [6, 28].

Since previous impressionistic descriptions suggest that mainstream AusE /l/ may be dark regardless of positional context [14, 15], we may expect darker onset /l/s for participants in our sample with exclusively monolingual English-speaking heritage. On the other hand, based on the aforementioned research from both AusE [19] and several UK English varieties [7, 9, 22], we also hypothesise that speakers in this study with Arabic and South Asian linguistic heritages will have clearer word-initial /l/s than speakers with solely English-speaking heritage. For participants with Chinese or Vietnamese language heritage, phonotactic evidence from their heritage language may lead us to predict that speakers with these linguistic heritages also have clearer onset /l/s, since Chinese [29] and Vietnamese [30] varieties do not allow coda /l/. Khmer does allow coda /l/ [31], but we do not have predictions as to how this may impact the realisation of onset /l/ among participants with Khmer-speaking heritage.

2. Methods

2.1. Materials

2.1.1. Data collection

This study uses a subset of data from the MAE-VoiS corpus [32]. All participants completed the same tasks, consisting of a picture-naming task with single words and short phrases elicited from images presented on a computer screen. Conversational data was also recorded but is not analysed here. Some participants (39 in total) were recorded remotely due to COVID-19 pandemic restrictions; the remainder were recorded in-person. Most participants were recorded at their schools, with a minority being recorded in other quiet settings such as their homes or local libraries.

2.1.2. Speakers

Participants were 148 high school students aged 15-18, recruited from different schools in the Sydney metropolitan area. They had completed all their schooling in Australia and all except 3 were also Australian-born; all participants were native speakers of AusE [32]. Participants were divided into 5 groups according to their heritage language background: monolingual English (22M, 33F), Arabic (15M, 5F), Chinese (including Mandarin and Cantonese) (9M, 11F), South-East Asian (Vietnamese and Khmer) (15M, 20F) and South Asian (including various Indo-Aryan and Dravidian languages) (12M, 7F). These groups were determined based on details of family language background provided by the participants and their guardian(s) collected through a demographic survey [32]. Most (but not all) of the speakers in the English heritage group were from Sydney's Northern Beaches, a largely homogeneous and monolingual area, whereas speakers from the other heritage language groups all resided in more culturally diverse areas in Sydney's west and south-west suburbs. Those in the Arabic, Chinese, South-East Asian and South Asian heritage language groups either spoke one of the languages in these respective groups at home or had exposure to that language through a parent/guardian. The non-English heritage language groups also reflect to varying degrees the communities in which these speakers reside—for example, most of those in the South-East Asian group are from the same area, as are those in the South Asian group, and so on, though there is not a one-to-one correspondence between language heritage group and area of residence [32]. The areas different speakers are recruited from vary demographically in other ways as well (such as general social class makeup), but these factors are beyond the scope of the current initial analysis.

2.1.3. /l/ items and positional contexts

Laterals analysed here were extracted from recordings of 36 single words and 11 short phrases. The laterals occurred in word-initial, medial, and final positional contexts. Word-medial /l/s, which we term *trochaic* (i.e. post-accentual—using terminology from [8, 9]) in the current study, are separated to assess whether they are acoustically more similar to onset or coda /l/s. We also categorised final /l/s occurring in an unstressed syllable (as in *bottle*) separately from those preceded by a stressed vowel (as in *doll*) since in the former context the lateral may form the syllable nucleus. This gives four /l/ positional contexts, shown in table 1, with each exemplified by one of the target words for clarity. 6438 /l/ tokens are analysed here, 3275 produced by female speakers and 3163 by male speakers.

Table 1. *Positional contexts of /l/ analysed.*

/l/ position	*N* items	Exemplar
initial	12	*leg*
final	10	*doll*
medial trochaic	9	*alligator*
syllabic /l/	6	*bottle*

2.2. Data preparation and analysis

2.2.1. Measuring /l/ darkness

F1 and F2 values were extracted at the midpoint of /l/ in an emuR database [33] using R [34] in RStudio [35]. This emuR database allowed inspection of all corpus items containing /l/ using spectrograms with segment boundaries which were automatically aligned using MAuS [36] with an AusE model. Outlier values were identified and hand-corrected where necessary, as were any notably large changes in formant

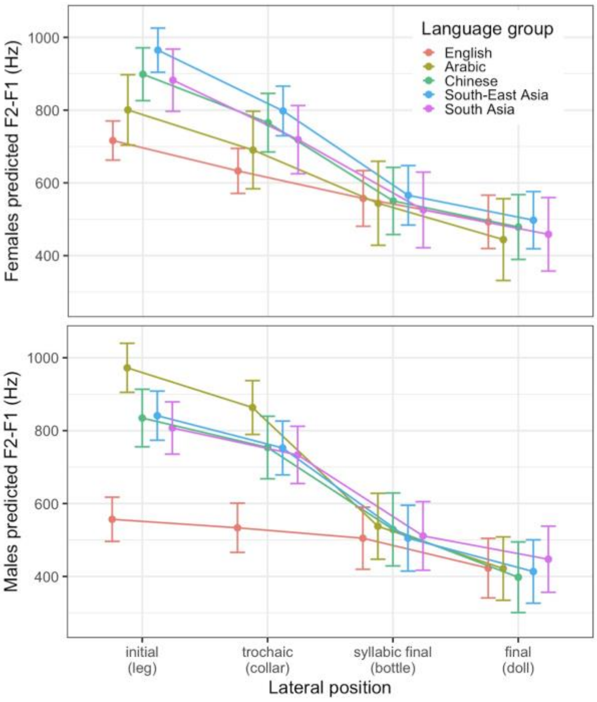


Figure 1: *Predicted midpoint /l/ F2–F1 by lateral position and language heritage group for female (top) and male (bottom) speakers.*

trajectories across the lateral that may have suggested errant formant tracks. After this, the difference between F2 and F1 (F2–F1) was calculated at the lateral midpoint.

Various /l/ darkness metrics have been used in previous studies [5, 19, 22]. Here we use raw F2–F1 (Hz) taken at the lateral midpoint, since both formants vary in the production of clear/dark /l/; this metric has been employed successfully in a range of previous studies [1, 3, 6]. The advantage of this approach is that it enables observation of how clear or dark laterals are in absolute terms [6].

We opt for using raw over normalised formant values because speakers in this corpus vary in the degree to which they exhibit short-front monophthong lowering, a change in progress in AusE [32, 37]. Today /æ/ forms the lower extremity of the vowel space for many, but not all, younger AusE speakers [32, 37]. As such, it may not be appropriate to normalise formant values given that the phonetically lowest vowel in the system varies between speakers. Since no normalisation is conducted beyond the effect of taking the difference between F2 and F1 [see 6], we analyse male and female speakers separately.

We turn briefly to vocalisation; some previous studies of AusE have opted to exclude vocalised tokens of /l/ [e.g. 19]. [38] suggests that the prevalence of vocalised /l/ varies regionally in AusE, being somewhat common in Sydney but less so than in cities like Adelaide. Vocalised /l/ is difficult to reliably distinguish from canonical dark /l/ via auditory and spectrographic evidence alone [17, 39]. We therefore have elected not to differentiate vocalised and canonical dark /l/ in this study, and include both.

2.2.2. *Data analysis*

We fitted linear mixed effects models to analyse lateral midpoint F2–F1 variability with the lme4 package [40]. Post-hoc analysis in the form of pairwise comparisons was conducted using the emmeans package, with Tukey corrections for *p*-values [41]. Separate models were fitted for the male and female speakers. Both modelled F2–F1 with fixed effects for language heritage group and lateral positional context and their interaction, with random intercepts included for word and participant. *P*-values for fixed effects and interactions were calculated using type III ANOVAs with the afex package [42].

3. Results

In both models, the interaction of language heritage group and /l/ positional context significantly affected F2–F1 (females: F(12, 3143.97) = 20.43, $p < .001$, males: F(12, 3033.29) = 54.64, $p < .001$). Figure 1 plots predicted variation in midpoint /l/ F2–F1 for female (top panel) and male (bottom panel) speakers, with higher F2–F1 values indicating clearer /l/s. It is apparent that the interaction affects /l/ darkness in both models: estimated F2–F1 varies between positional contexts for all language groups, but to different degrees. The subsequent sections describe results from pairwise comparisons which show significant differences between /l/ positional contexts and heritage language groups.

3.1. Female speakers

3.1.1. */l/ darkness between positional contexts—females*

The top panel of figure 1 shows that generally, the female speakers in all language groups produce a positional distinction between word-initial and word-final /l/. Results from pairwise comparisons confirm these observations: all groups have significantly higher F2–F1 in word-initial /l/s (as in *leg*) than in word-final (*doll*, all $p < .0001$) or syllabic final /l/s (*bottle*, all $p < .001$). All groups also had clearer /l/s in medial trochaic contexts (e.g. *alligator*) than in word-final contexts (all $p < .007$), and all apart from the English heritage group had clearer trochaic /l/s than syllabic final /l/s as well (English $p = .331$; all others $p < .048$). Every group also had clearer initial /l/s than medial trochaic /l/s (all $p < .029$). None of the groups differed significantly in final vs. syllabic final /l/ contexts.

3.1.2. */l/ darkness between diversity groups—females*

For the female speakers, /l/ darkness appears to vary between heritage language groups in different positional contexts, with the English and Arabic heritage groups producing darker /l/s in certain contexts than the other groups. Post-hoc analysis confirms this: the English heritage group had significantly darker word-initial /l/s than the Chinese ($p < .0001$), South-East Asian ($p < .0001$), or South Asian ($p = .0007$) heritage groups, and the Arabic heritage group also had significantly darker /l/s than the South-East Asian heritage group in this context ($p = .0091$). Furthermore, the English heritage group had significantly darker trochaic /l/s than the Chinese and South-East Asian heritage groups (both $p < .004$). None of the heritage language groups differed significantly from one another in word-final or syllabic final /l/ contexts.

3.2. Male speakers

3.2.1. */l/ darkness between positional contexts—males*

As can be seen in the bottom panel of figure 1, the male speakers in all groups seem to produce clearer word-initial /l/s and darker word-final /l/s. Again, results from pairwise comparisons confirm this. For all groups word-initial /l/ (e.g. *leg*) had significantly higher F2–F1 than final /l/ (e.g. *doll*: all $p < .0126$). Trochaic /l/ (e.g. *alligator*) was significantly clearer than final /l/ for all groups except the English heritage group (English $p = .0811$; all other groups $p < .0001$). The Arabic, Chinese, South-East Asian and South Asian heritage groups also showed significantly clearer initial /l/s than those in both syllabic final (all $p < .0001$) and medial trochaic contexts (all $p < .048$), but this was not true for the English heritage group for either of these contrasts; these same four groups also produced clearer trochaic /l/s than syllabic final /l/s (all groups $p < .0002$), whereas again the English heritage group did not. None of the groups differed between word-final versus syllabic final /l/ contexts.

3.2.2. /l/ darkness between diversity groups—males

As with the female speakers, there are noticeable differences in /l/ darkness between language heritage groups of male speakers, most obviously in word-initial and medial trochaic contexts. The Arabic, Chinese, South-East Asian and South Asian heritage groups all had significantly clearer /l/s than the English heritage group in initial (all $p < .0001$) and trochaic contexts (all $p < .0001$).

In addition to this, the Arabic heritage group had significantly clearer word-initial /l/s than the Chinese, South-East Asian and South-Asian heritage groups (all $p < .0027$), and clearer trochaic /l/s than the South Asian and South-East Asian groups (both $p < .0301$). As with the females, none of the male language heritage groups differed significantly from each other in their production of word-final or syllabic final /l/s.

4. Discussion

Broad patterns are observable in the results, both for female and male speakers. In general, onset /l/s were clearer than coda /l/s for all groups, regardless of gender or heritage language background. We observed that trochaic /l/s (e.g. *alligator*) were significantly clearer than word-final laterals, exhibiting characteristics more similar to onset /l/s than coda /l/s, as has been found in some previous work [6, 28]. However, we also saw that some groups had significantly clearer word-initial /l/s than medial trochaic /l/s, which may be attributed to coarticulatory factors. As discussed earlier, clearer /l/s are known to be more susceptible to coarticulatory influence from surrounding vowels than darker /l/s [3, 12, 13]; therefore the relatively darker /l/s we observed word-medially may have been realised in this manner as they experience coarticulatory influence from both preceding and following segments, whereas word-initial /l/s do not. Future research should investigate this further.

The results also make clear the extent to which language heritage may impact lateral realisation. For females, the English (and to an extent Arabic) heritage speakers produced darker word-initial and trochaic than those in the Chinese, South-East Asian and South Asian heritage groups. Among the males, the English heritage speakers also produced darker onset /l/s than the Chinese, South-East Asian and South Asian heritage group speakers, but here the Arabic heritage speakers can be characterised as producing "hyper-clear" onset /l/s, with F2–F1 values higher than all other groups. Previous work has suggested that clearer onset /l/s may be a feature inherited from heritage languages [e.g. 9, 22, 23, 24]. In many of the participants' heritage languages—including Arabic [19, 22], Punjabi [24], Hindi [43], Cantonese [44], and even common L2 varieties like Indian English [45]—/l/ is described as acoustically clear. On the other hand, the model predictions of word-initial /l/ F2–F1 for both the female and male English heritage speakers (approx. 750 Hz and 620 Hz respectively) are similar those reported in Manchester [6, 11] and Glasgow [46], where onset /l/ is described as dark. Our results therefore confirm previous descriptions of AusE /l/ as being acoustically dark in all contexts for participants with monolingual English-speaking heritage [14, 15, 20], but suggest that this characterisation is limited in the context of increased diversity in the speech community—as [19] also shows.

However, the Arabic heritage male and female speakers appear to realise initial and medial /l/s differently, with males having very clear /l/s relative to other groups of the same gender, and females somewhat darker than most. We do not directly compare male and female speakers in this analysis, and we note that the number of Arabic heritage females ($n = 5$) is quite low, but this observation complicates the idea that differences in /l/ darkness are attributable solely to heritage language transfer. One of the key findings in [19] was that Lebanese heritage speakers (both male and female) with more densely Lebanese social networks produced clearer initial /l/s than those with fewer Lebanese contacts. It may be that the difference between males and females of Arabic heritage in our sample reflects the extent to which individual speakers are integrated into their heritage language communities. More research is needed to investigate this hypothesis.

Lastly, we turn to some limitations of the current analysis and directions for future study. This study uses citation form data, so investigation of conversational /l/ productions is required. Our use of lateral F2–F1 midpoint as a metric of /l/ darkness also precludes examination of formant trajectories transitioning into and out of the lateral. A future analysis could use dynamic formant analyses to better assess the effect of coarticulation from adjacent segments on /l/ darkness.

5. Conclusion

This study reported on variation in /l/ darkness in AusE, showing that speakers produce a positional distinction between clearer initial and darker final /l/s, corroborating previous research [19, 20] and extending these analyses by incorporating /l/s elicited word-medially. Furthermore, we showed that speakers with different heritage language background varied in their production of initial and medial /l/s: speakers with a non-English language heritage generally produced clear /l/s in these contexts, whereas speakers with English-only backgrounds produced much darker onset /l/s. We have also outlined directions for future research, including closer consideration of coarticulatory factors and incorporation of dynamic analyses.

6. Acknowledgements

This research was supported by a Macquarie University Research Excellence Scholarship to the first author, and ARC Future Fellowship Grant FT180100462 to the third author. We also wish to thank members of the MQ Phonetics lab for their insightful contributions to this analysis.

7. References

[1] Sproat, R., and Fujimura, O., "Allophonic variation in English /l/ and its implications for phonetic implementation", J. Phon., 21(3):291-311, 1993.

[2] Ladefoged, P., and Maddieson, I., "The sounds of the world's languages", Blackwell, 1996.

[3] Recasens, D., "A cross-language acoustic study of initial and final allophones of /l/", Speech Commun., 54(3):368-383, 2012.

[4] Recasens, D., and Espinosa, A., "Articulatory, positional and coarticulatory characteristics for clear /l/ and dark /l/: evidence from two Catalan dialects", J. Int. Phon. Assoc., 35(1):1-25, 2005.

[5] Kirkham, S., Turton, D., and Leeman, A., "A typology of laterals in twelve English dialects", J. Acoust. Soc. Am., 148(1):EL72–76, 2020.

[6] Kirkham, S., Nance, C., Littlewood, B., Lightfoot, K., and Groarke, E., "Dialect variation in formant dynamics: the acoustics of lateral and vowel sequences in Manchester and Liverpool English", J. Acoust. Soc. Am., 145(2):784-794, 2019.

[7] Stuart-Smith, J., "Glasgow: accent and voice quality", in P. Foulkes and G. Docherty [Eds], Urban voices: accent studies in the British Isles, 203-222, Arnold, 1999.

[8] Carter, P., and Local, J., "F2 variation in Newcastle and Leeds English liquid systems," J. Int. Phon. Assoc., 37(2):183-199, 2007.

[9] Kirkham, S., "Ethnicity and phonetic variation in Sheffield English liquids", J. Int. Phon. Assoc., 47(1):17-35, 2017.

[10] Stuart-Smith, J., Timmins., C., and Alam, F., "Hybridity and ethnic accents", in F. Gregersen, J. Parrot, and P. Quist [Eds], Language variation: European perspectives III, 43-59, John Benjamins, 2011.

[11] Turton, D., and Baranowski, M., "The sociolinguistics of /l/ in Manchester", Ling. Van., 7(1):1-11, 2021.

[12] Simonet, J., "An acoustic study of coarticulatory resistance in "dark" and "light" laterals", J. Phon. 52:138-151, 2015.

[13] Farnetani, E., and Recasens, D., "Coarticulation and connected speech processes", in W. J. Hardcastle, J. Laver, and F. E. Gibbon [Eds], The handbook of phonetic sciences: second edition, 316-352, Wiley, 2010.

[14] Cox, F., and Fletcher, J., "Australian English pronunciation and transcription", 2nd edition, Cambridge University Press, 2017.

[15] Wells, J. C., "The southern hemisphere: Australia", in J. C. Wells [Ed], Accents of English 3: beyond the British Isles, 592-604, Cambridge University Press, 1982.

[16] Loakes, D., Clothier, J., Hajek, J., and Fletcher, J., "An investigation of the /el/-/æl/ merger in Australian English: a pilot study on production and perception in south-west Victoria", Aus. J. Ling., 34(4):436-452, 2014.

[17] Szalay, T., Benders, T., Cox, F., Palethorpe, S., and Proctor, M., "Spectral contrast reduction in Australian English /l/-final rimes", J. Acoust. Soc. Am., 149(2):1183-1197, 2021.

[18] Harrington, J., and Schiel, F., "/u/-fronting and agent-based modelling: the relationship between the origin and spread of sound change", Language 93(2):414-445, 2017.

[19] Clothier, J., "A sociophonetic analysis of /l/ darkness and Lebanese Australian ethnic identity in Australian English", in S. Calhoun, P. Escudero, M. Tabain, and P. Warren [Eds], Proc. 19th ICPhS, Melbourne, 1888-1892, Australasian Speech Science and Technology Association Inc., 2019.

[20] Schmidt, P., Diskin-Holdaway, C., and Loakes, D., "New insights into /el/-/æl/ merging in Australian English", Aus. J. Ling., 41(1):66-95, 2021.

[21] Barlow, J. A., Branson, P. E., and Nip, I. S. B., "Phonetic equivalence in the acquisition of /l/ by Spanish-English bilingual children", Biling.: Lang. Cog., 16(1):68-85, 2013.

[22] Khattab, G., "Acquisition of Lebanese Arabic and Yorkshire English /l/ by bilingual and monolingual children: a comparative spectrographic study", in Z. M. Hassan and B. Heselwood [Eds], Instrumental studies in Arabic phonetics, 325-354, John Benjamins, 2011.

[23] Kirkham, S., and McCarthy, K. M., "Acquiring allophonic structure and phonetic detail in a bilingual community: the production of laterals by Sylheti-English bilingual children", Int. J. Biling., 25(3):531-547, 2021.

[24] Kirkham, S., and Zara, M., "Intergenerational transmission of laterals in Punjabi-English heritage bilinguals", in R. Rao [Eds], The phonetics and phonology of heritage languages, 129-146, Cambridge University Press, 2024.

[25] Hall, K., and Bucholtz, M., "From mulatta to mestiza: language and the reshaping of ethnic identity", in K. Hall and M. Bucholtz [Eds], Gender articulated: language and the socially constructed self, 351-374, Routledge, 1995.

[26] Willoughby, L., and Manns, H., "Introducing Australian English", in L. Willoughby and H. Manns [Eds], Australian English reimagined: structure, features and developments, 1-12, Routledge, 2019.

[27] Australian Bureau of Statistics, "Greater Sydney: 2021 Census all persons QuickStats", 2021. Online: https://abs.gov.au/census/find-census-data/quickstats/2021/1GSYD, accessed 14 Jun 2024.

[28] Lee-Kim, S., Davidson, L., and Hwang, S., "Morphological effects on the darkness of English intervocalic /l/", Lab. Phon., 4(2):475-511, 2013.

[29] He, Y., "Production of English syllable final /l/ by Mandarin Chinese speakers", J. Lang. Teach. Res., 5(4):742-750, 2014.

[30] Kirby, J., "Vietnamese (Hanoi Vietnamese)", J. Int. Assoc. Phon., 41(3):381-392, 2011.

[31] Nacaskul, K., "The syllabic and morphological structure of Cambodian words", Mon-Khmer Stud. J., 7, 183-200, 1978.

[32] Cox, F., and Penney, J., "Multicultural Australian English: the new voice of Sydney", Aust. J. Ling., advance online publication, 2024.

[33] Jochim, M., Winkelmann, R., Jaensch, K., Cassidy, S., and Harrington, J., "emuR: main package of the EMU Speech Database Management System", R package version 2.5.0, 2024.

[34] R Core Team, "R: a language and environment for statistical computing", R Foundation for Statistical Computing, 2024.

[35] Posit Team, "RStudio: integrated development environment for R", Posit Software PBC, 2024.

[36] Schiel, F., Draxler, C., and Harrington, J., "Phonetic segmentation and labelling using the MAUS technique", New Tools and Methods for Very-Large-Scale Phonetics Research Workshop, 1-4, 2011.

[37] Cox, F., Penney, J., and Palethorpe, S., "Australian English monophthong change across 50 years: dynamic vs. static measures", Languages, 9(3):1-35, 2024.

[38] Horvath, B. M., and Horvath, R. J., "A multilocality study of a sound change in progress: the case of /l/ vocalization in New Zealand and Australian English", Lang. Var. Change, 13(1):37-57, 2001.

[39] Hall-Lew, L., and Fix, S., "Perceptual coding reliability of (L)-vocalization in casual speech data", Lingua, 122(7):794-809, 2012.

[40] Bates, D., Mächler, M., Bolker, B., and Walker, S., "Fitting linear mixed effects models using lme4", J. Stat. Soft., 67(1):1-48, 2015.

[41] Lenth, R., "emmeans: estimated marginal means, aka least-square means", R package version 1.10.2, 2024.

[42] Singmann, H., Bolker, B., Westfall, J., Aust, F., and Ben-Schachar, M., "afex: analysis of factorial experiments", R package version 1.3-1, 2024.

[43] Shaktawat, D., "The effect of Indian contact and Glaswegian contact on the phonetic backwards transfer of Glaswegian English (L2) on Hindi and Indian English (L1)", Languages, 9(4):118-145, 2024.

[44] Chan, A. Y. W. and Li, D. C. S., "English and Cantonese phonology in contrast: explaining Cantonese ESL learners' English pronunciation problems", Lang. Cul. Curric. 13(1):67-85, 2000.

[45] Gargesh, R., "Indian English: Phonology", in B. Schneider and E. W. Kortmann [Eds], A handbook of varieties of English, 1st edition, 992-1002, Mouton de Gruyter, 2008.

[46] Macdonald, R., and Stuart-Smith, J., "Coarticulation guides sound change: an acoustic-phonetic study of real time change in word-initial /l/ over four decades of Glaswegian", in F. Kleber and T. Rathke [Eds], Speech dynamics: synchronic variation and diachronic change, De Gruyter Mouton, in press.

A Doll on the Dole: Prelateral Length Mergers in Australian English

Joshua van de Ven[1], *Brett Baker*[1], *Rikke Bundgaard-Nielsen*[1], *Yizhou Wang*[1]

[1]Department of Linguistics, University of Melbourne, Australia.

jvandevenlinguist@gmail.com, bjbaker@unimelb.edu.au, rikkieb@unimelb.edu.au, yizwang3@unimelb.edu.au

Abstract

The effects of prelateral coarticulation on vowels in Australian English sometimes result in the reduction of spectral contrastiveness in vowel pairs. As a result, some speakers may struggle to distinguish minimal pairs that differ only in these vowels prelaterally, for instance realising both *doll* and *dole* as /dəʉl/. This pilot study investigates the effect of this context-induced merger by testing the accuracy and reaction time of 10 young Australian English speakers from Melbourne in an AXB discrimination and categorisation task. Participants struggled to distinguish some prelaterally merged vowel pairs more than others. There were also significant differences between the categorisation and discrimination results for front and back short/long vowel pairs.

Index Terms: prelateral coarticulation, lateral-final rimes, perception, Australian English.

1. Introduction

It is well known that changes in the perception of coarticulatory effects on phonemes can lead to sound change [1-5]. In English, prelateral coarticulation can lead to a reduction in acoustic and perceptual contrasts in some vowels [5-10]. This has led to a number of changes in English prelateral vowel contrasts, including the reduction of spectral contrast in the tense-lax FLEECE-KIT vowel pair in Standard Southern British English [11], the reduction of the POOL-PULL-POLE contrast in Ohio English [9], and the /el/~/æl/ merger in New Zealand and Melbourne English [12-16], which is of particular interest in relation to the participant population of the present study. In Australian English (AusE) specifically, prelateral coarticulation has been particularly important in triggering a number of sound changes besides the /el/~/æl/ merger [10], including the developing /ɐlC/~/ɔlC/ merger [17] and the reduction of spectral contrast in some prelateral vowel pairs [6], [18].

The AusE vowel inventory consists of 18 contrastive vowels, varying in both spectral and durational quality [19]. Within this inventory there are three vowel pairs which exhibit substantial allophonic overlap in prelateral contexts – /ʉː-ʊ/, /æɔ-æ/, /əʉ-ɔ/ (GOOSE-FOOT, MOUTH-TRAP, GOAT-CLOTH) [5], [8], [20]. These contrasts are spectrally reduced before /l/ such that there is little difference in vowel quality, though the length distinction between them may be maintained [6], [18], [21]. The length contrast is not always perceptible, however, as a result of ascribing the lengthening of the vowel to a coarticulatory effect of the subsequent /l/ on the preceding vowel. This leads to difficulties in disambiguating the vowel from the following /l/ [5]. As a result, /ʉː-ʊ/, /æɔ-æ/, and /əʉ-ɔ/ may merge prelaterally, leading to the homophonous pronunciation of pairs like *doll* versus *dole* as /dəʉl/ [6], [21].

A number of studies have investigated this context-induced merger of AusE /ʉː-ʊ/, /æɔ-æ/, and /əʉ-ɔ/. Szalay et al. [20] tested whether speakers of AusE could successfully transcribe words with these vowel contrasts using English orthography (i.e. *fool-full, howl-Hal,* etc.). They found that speakers generally struggled to correctly transcribe minimal pairs featuring vowel contrasts that they did not themselves produce. Szalay et al. [5], [8] similarly concluded that reduced perceptual contrast in these prelateral vowels may lead to a vowel merger in AusE, but did not find perception to be skewed towards a particular phonemic category in each pair (i.e. a skew towards interpreting an ambiguous vowel as a short or long vowel). These findings are informative, but Szalay et al.'s [5], [20] results were potentially affected by word-frequency effects [20] and the mixed use of real and nonce words [5], both of which are known to have confounding effects on speaker perception [22].

Design limitations aside, Szalay et al.'s findings suggest that prelateral context has a strong effect on the perception of /ʉː-ʊ/, /æɔ-æ/, and /əʉ-ɔ/ in AusE. Given that changes in perception as a result of coarticulatory effects are often a precursor to sound change [1-4], further investigation into the perception of these prelateral contrasts may provide useful insight into emerging sound changes in Australian English. To that end, this paper investigated two research questions through an AXB discrimination task and a categorisation task:

1. To what extent is a merger of the vowel pairs /ʉː-ʊ/, /æɔ-æ/, and /əʉ-ɔ/ underway prelaterally in the AusE of young Melbournians?
2. Is there a preference for interpreting vowels in each of these pairs as long or short where their length is ambiguous prelaterally?

2. Method

The present study was co-designed by the students in the 2024 Semester 1 subject LING40009 Seminars in Descriptive Linguistics at the School of Linguistics at the University of Melbourne. Data collection was undertaken by the first author, Geordie Kidd, Katie Carpenter, Felix Kimber, and Lizzie Kelly.

2.1 Participants

Ten native speakers of AusE (five female; four male; one other) participated in this pilot study. The participants ranged in age from 17-27 (*M* age = 21.5). All participants were born and raised in Melbourne to monolingual English-speaking parents and were not linguistically trained. None reported any reading, hearing, or speaking disorders. The ten participants were recruited through the first author's personal network. They were not paid for their participation. The study was approved by the University of Melbourne's human ethics committee (reference number: 2024-29383-52826-3).

2.2 Materials

Seven nonce words of the form *snVl* were used for both the discrimination and the categorisation experiments outlined below. Nonce words were chosen to avoid the potentially confounding effect of lexical frequency in aiding perception. The nonce words included the three short/long vowel pairs of interest here (/snæl-snæɔl, snɔl-snəʉl, snʊl-snʉːl/) as well as a control pair (/snɪl-snəʉl/). The stimulus material was produced by three Melbourne-raised native male speakers of AusE aged 22-23. The speakers were recorded in a quiet location. They were familiarised with the intended target nonce words through the presentation of real rhyme words and instructed to produce the target words in a clear, casual manner, as if speaking to a friend. The estimated average formant frequencies and rhyme lengths across all three speakers for the stimulus material are presented in *Table 1* below. Formant frequencies were estimated for the midpoint of the vowel. Rhyme lengths were measured given the difficulty of segmenting vowel + lateral coda sequences.

Table 1. *Estimated average formant frequency and rhyme length data across all three stimulus speakers for stimulus material used in the experiment.*

Word	*F1 (Hz)*	*F2 (Hz)*	*Rhyme length (ms)*
/snæl/	738	1575	298
/snæɔl/	645	1746	431
/snɔl/	559	1112	311
/snəʉl/	553	1119	465
/snʊl/	473	1112	247
/snʉːl/	448	1020	391
/snɪl/	503	1672	277

2.3 Procedures

Participants completed a categorisation task and an AXB discrimination task. This was done in order to determine whether participants are able to distinguish these prelateral vowel pairs when heard together (in the AXB discrimination task) and in isolation (in the categorisation task) to ascertain the degree of merger that may be underway, in line with research question 1. The order of perception tasks was randomised such that half of the participants began with the AXB task and half began with the categorisation task. Both tasks were created and run using Praat version 6.4.11 [23]. All participants completed the two tasks on a laptop or desktop computer in a quiet room using good quality headphones. Per participant randomisation was applied for all tasks.

2.3.1 Categorisation task

In the categorisation task, participants were shown a screen with seven real English words in English orthography. They then heard one of the seven *snVl* nonce words used in the AXB task and were asked to choose the corresponding English word which best rhymed with the nonce word. They were then asked to rate how well they thought the word they selected rhymed with the heard nonce word on a five-point Likert scale as a goodness-of-fit measure. The seven real English words used as rhymes for the nonce words were *pal, foul, doll, pole, bull, pool*, and *pill*. Potentially confounding effects of frequency [22] were reduced where possible, with a mean 16,278 parts per million in COCA [24] (range 9,198-42,491; SD 10,983). Words were selected which began with a labial consonant to minimise vowel-to-consonant coarticulation [25-26]. The exception to this was *doll*, which was chosen because it is more regularly pronounced with the short [ɔ] than the labial-initial *poll* (which is more often rendered [pəʉl]) and has a similar frequency to the other words selected for the study (10,794 for *doll*, cf. 31,129 for *poll*).

Participants completed a total of 63 trials, hearing each recording of each nonce word three times per speaker (7 nonce words x 3 speakers x 3 repetitions = 63). They completed a practice task with unrelated *snVl* nonce words prior to the experiment to ensure they understood the task. The participants were familiarised with the target words via a recording by a male 22-year-old native AusE speaker from Melbourne producing the seven English rhyming words. This was to ensure participants understood the pronunciation of each of the English words on screen before beginning the categorisation task (i.e. that they did not initially misread the categorisation target words). Recordings of this speaker were not used elsewhere in the experimental design.

2.3.2 Discrimination task

In the AXB discrimination task, participants heard a sequence of three stimuli (A, X, B) where the middle stimulus was phonologically identical to either the first (A) or third stimulus (B). Each stimulus was produced by a different one of the three stimulus speakers such that participants were forced to identify phonological categories to complete the task, rather than relying on speaker-specific voice patterns for identification. The ordering of speakers was randomised. An interstimulus length of 300ms and intergroup pause of 500ms were used.

Participants completed a total of 96 trials examining four word-pairs – the three target pairs (/snæl-snæɔl, snɔl-snəʉl, snʊl-snʉːl/) and a control (/snɪl-snəʉl/). Each pair was presented 24 times (2 target positions x 2 X values x 6 speaker configurations = 24). Participants completed a practice task with unrelated *snVl* nonce words prior to the main experiment to ensure they understood the task.

2.4 Data analysis

The discrimination task yielded data in the form of discrimination accuracy and reaction time (RT), collected automatically from Praat. The categorisation task yielded perceptual categorisation accuracy data and goodness-of-fit ratings for each categorised *snVl* word to the chosen English rhyme, collected automatically from Praat. Repeated measures ANOVAs were conducted for the AXB task results using JASP [27]. Accuracy and goodness ratings in the categorisation task were compared in a confusion matrix.

3. Results

3.1. Results of categorisation task

The results for the categorisation task are presented in the confusion matrix in *Table 2* below. Percentage correct target categorisations are presented in bold on the diagonal with the mean goodness-of-fit rating in parentheses.

Almost all non-target categorisations were of the other length value in the target pair (e.g. incorrectly categorising /snʊl/ as rhyming with *pool* rather than *bull*, etc.). The exceptions were /snɔl/ and /snəʉl/, notably the least accurate pair, which were also incorrectly categorised as rhyming with *bull* (7%) and *foul* (2%) respectively. There are clear asymmetries in accuracy within each short-long vowel pair. For the front pair /snæl-snæɔl/, participants were more accurate identifying the short vowel. The opposite is true for the two back vowel pairs, /snɔl-snəʉl/ and /snʊl-snʉːl/, where participants were more accurate identifying the long vowel.

Table 2. *Confusion matrix giving percentage accuracy for each response in the categorisation task. Correct responses are bolded, with mean goodness ratings shown in brackets. Responses which only occurred once are not included.*

		Response						
		/snɪl/ (control)	/snæl/	/snæɔl/	/snɔl/	/snɔʉl/	/snʊl/	/snʉːl/
Stimulus	/snɪl/ (control)	**99** **(4.22)**						
	/snæl/		**72** **(3.57)**	26 (3.17)				
	/snæɔl/		41 (3.14)	**58** **(3.67)**				
	/snɔl/				**42** **(4.05)**	47 (3.48)	7 (1.67)	
	/snɔʉl/			2 (2.50)	20 (3.89)	**76** **(3.96)**		
	/snʊl/						**69** **(3.26)**	30 (3.48)
	/snʉːl/						22 (3.30)	**77** **(3.83)**

3.2 Results of AXB discrimination task

The discrimination accuracy level of each vowel contrast is presented in *Figure 1*. A repeated measures ANOVA was performed to assess the effect of vowel contrast on discrimination accuracy. Since Mauchly's test of sphericity was significant, Greenhouse-Geisser correction was used. This showed that discrimination accuracy differed significantly $F(2.402,9)=11.692$, $p<0.001$ across the four contrasts. Post hoc testing using Bonferroni correction showed that the control pair is more accurately discriminated than the /snæl-snæɔl/ pair (M difference = −0.161, p = 0.003) and the /snʊl-snʉːl/ pair (M difference = −0.108, p = 0.013). Differences between the control and the /snɔl-snəʉl/ pair were not significant.

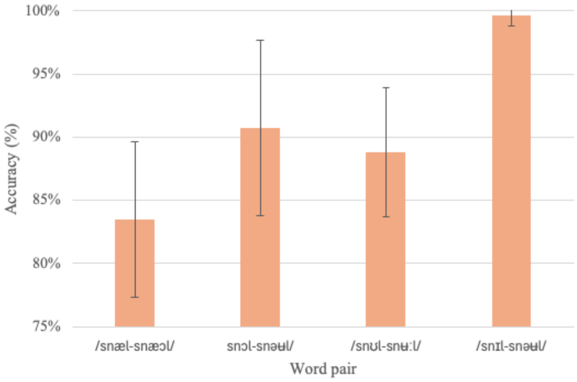


Figure 1. *Descriptive plot showing the effect of word pair on accuracy. Confidence interval of 95% shown.*

There was also a statistically significant difference in mean reaction time for the four contrasts $F(2.147,9)=4.075$, $p=0.031$, shown in *Figure 2*. Bonferroni corrected post-hoc comparisons showed that participants were more quickly able to distinguish the control pair than the /snɔl-snəʉl/ pair (M difference = 204ms, p = 0.022) and the /snʊl-snʉːl/ pair (M difference = 347ms, p = 0.023). Differences between the control and the /snæl-snæɔl/ pair were not significant.

A third repeated measures ANOVA was performed to compare the effect of the length of the X vowel on accuracy within each target pair (e.g. for the /snæl-snæɔl/ pair, comparing whether participants were more accurate if X was /snæl/ or /snæɔl/, etc.), following Thomas and Hardy [16]. Since Mauchly's test of sphericity was significant, Huynh-Feldt correction was used. Length of the X vowel had a significant effect on discriminating the /snæl-snæɔl/ vowel pair, as shown in *Figure 3*, but not /snɔl-snəʉl/ or /snʊl-snʉːl/ (M difference = –0.168, p < .001). For the /snæl-snæɔl/ pair, participants were significantly more accurate where X was a long vowel (/snæɔl/). This is interesting in light of the categorisation task, where participants were more accurate in categorising /snæl/ than /snæɔl/.

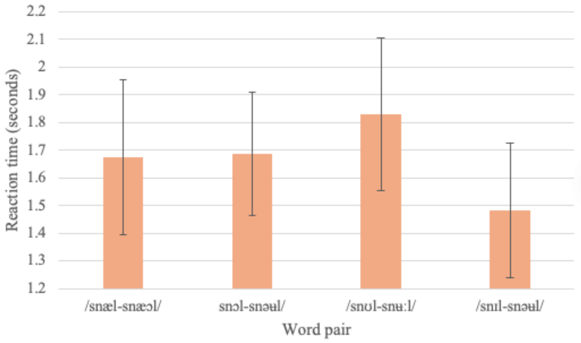


Figure 2. *Descriptive plot showing the effect of word pair on reaction time. Confidence interval of 95% shown.*

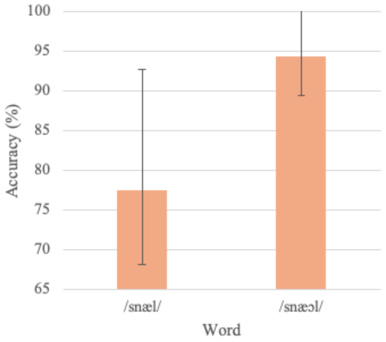


Figure 3. *Descriptive plot showing the effect of length of the X vowel on accuracy for the* /snæl-snæɔl/ *pair. Confidence interval of 95% shown.*

A further repeated measures ANOVA was conducted to see if there was a statistically significant effect of ordering on accuracy (e.g. if participants were more accurate at discriminating a long vowel if they heard a long vowel first). No statistically significant effect was found.

4. Discussion

The present study investigated the emerging context-dependent merger of the short/long vowel pairs /ʉː-ʊ/, /æɔ-æ/, and /əʉ-ɔ/ in the variety of Australian English spoken by young people in Melbourne.

The categorisation task presented in Section 3.1 indicates that participants find it difficult to accurately categorise /ʉː-ʊ/, /æɔ-æ/, and /əʉ-ɔ/ vowels prelaterally. For the front /æ-æɔ/ pair, participants were more accurate in categorising the short /æ/ vowel. The opposite was true for the back /ʉː-ʊ/ and /əʉ-ɔ/ pairs, where participants were more accurate in categorising the long /ʉː/ and /əʉ/ vowels.

The AXB discrimination task results presented in Section 3.2 indicate that participants also find it difficult to accurately discriminate /ʉː-ʊ/ and /æɔ-æ/ vowels prelaterally, consistent with the results of the categorisation task. The discrimination task also indicated that participants were more accurate where the target vowel was long for the /snæl-snæɔl/ pair, which is the opposite finding to the categorisation task. No statistically significant effect of length on accuracy was found for the /snɔl-snəʉl/ or /snʊl-snʉːl/ vowels.

In most psycholinguistic experiments, reaction times (RTs) are assumed to indicate how difficult participants find a task to complete. Faster RTs indicate that a task is easy; longer RTs indicate that a task is difficult. RTs for the /ʉː-ʊ/, /əʉ-ɔ/, and /æɔ-æ/ pairs in the discrimination task were all longer (significance levels reached only for the first two pairs) when compared to the control, suggesting that participants find it difficult to discriminate these vowels prelaterally.

Together, these results suggest that a complete merger has not occurred for prelateral /ʉː-ʊ/, /æɔ-æ/, and /əʉ-ɔ/ in the AusE of all young people in Melbourne. Nevertheless, the decreased accuracy of discrimination and categorisation in these pairs compared with the control suggests that a partial merger may be underway, answering research question 1. Specifically, comparatively high accuracy in the discrimination task compared to the categorisation task suggests that participants can still hear a difference in the target vowels when presented together, as would be expected if no merger has occurred. The lower accuracy of the categorisation task is difficult to explain, however, unless these pairs have undergone some degree of merger.

Despite the methodological differences (real versus nonce word stimuli), the results here are consistent with the findings of Szalay et al. [5], [8], [20]. We find evidence that speakers struggle to discriminate differences in AusE length pairs in prelateral contexts even when accounting for the confounding factors of frequency and mixed use of nonce and real words in the experimental design, giving further credence to Szalay et al.'s previous findings.

In relation to research question 2, our data may suggest that participants interpret an ambiguous vowel as long prelaterally. The participants in the present study tend to be more accurate in categorising the long back /ʉː/ and /əʉ/ vowels and discriminating the long front /æɔ/ vowel. These findings do not necessarily support Szalay et al.'s [5] claim that there is no preference for a particular length where it is difficult to distinguish a vowel in prelateral contexts. This potential preference for a long vowel may be a result of partial merger, where the merged vowel in the speech of the participants is closer to a long vowel target, as we might expect given the anecdotal preference for pronouncing *doll* and *poll* as /Cəʉl/. As a result, participants can more accurately categorise the long vowel because it more closely matches their own merged vowel, and are less accurate in categorising the short vowel, which does not occur as often in their speech. The results here may therefore resemble other cases of partial merger, such as the /el/~/æl/ merger in New Zealand English [16], as well as Szalay et al.'s previous findings [20]. Production data from participants would be required to investigate this further.

This does not explain why participants were more accurate in categorising the short /snæl/ vowel in the categorisation task, however. This could potentially be explained by appealing to hypercorrection in the sense of Ohala [4] as a result of compensation for /l/ vocalisation. /l/ vocalisation is a common process in English which results in the realisation of /l/ as a high back vowel /ʊ/ and is known to affect speech sound perception [8], [10], [31-33]. Participants may thus be interpreting the back vowel in /snæɔl/ as a vocalised /l/ (e.g. [snæɔʊ]) and correspondingly overcorrect their perception to /snæl/. Given that /snæl-snæɔl/ begins with a front vowel target, there is a greater distinction between the target nucleus and the lateral when compared to the /snɔl-snəʉl/ or /snʊl-snʉːl/ back vowel pairs. As a result, it is possible that the diphthong is camouflaged by the /æ/ to /l/ trajectory such that participants are more susceptible to overcorrect for the effects of /l/ coarticulation in this front vowel than for the back vowels.

It is also possible that what appears to be a preference for interpreting an ambiguous vowel as long could in fact be explained by neighbourhood effects – e.g. participants may have been more accurate in discriminating the long /snʉːl/ nonce word than the short /snʊl/ because more words in English end in /-ʉːl/ (*fool, tool, cool, drool, stool,* etc.) than /-ʊl/ (*full, bull,* etc.). Possible neighbourhood effects were not specifically accounted for in the experimental design here, and thus it is difficult to determine with certainty what is motivating this possible preference for long vowels. Future experimental work would benefit from including both production data and accounting for neighbourhood effects to better explore this behaviour.

A further shortcoming in our experimental design was the decision to avoid minimal pairs in the categorisation task target rhymes (e.g. *doll* vs *pole* for the /snɔl-snəʉl/ pair rather than *doll* vs *dole*), as this increased the difficulty of the task such that accuracy was substantially lower than has been found in other studies. Half of the participants had an accuracy of lower than 50% for /snɔl/. Participants therefore found distinguishing the /snɔl-snəʉl/ pair particularly difficult. This may be a consequence of both *doll* and *pole* being realised as /Cəʉl/ for the participants, with the phonological shape of *doll* being more ambiguous as a result of the possible merger that is the focus of this paper. As a result, participants perhaps prefer the more familiar word *pole*, resulting in a much lower than 50% accuracy for /snɔl/. Future experimentation would thus benefit from deliberately incorporating minimal pairs in the experimental design – e.g. asking participants to rhyme with either *doll* or *dole*, rather than *doll* and *pole*.

5. Conclusions

Coarticulatory effects of /l/ on vowel length may be leading to a merger of the long-short /ʉː-ʊ/, /æɔ-æ/, and /əʉ-ɔ/vowel pairs prelaterally in AusE. In particular, speakers of AusE may have a general preference for perceiving these vowels as long where their length is ambiguous prelaterally. This may be a result of production influencing perception, in line with Szalay et al.'s [20] findings. This preference may be confounded by the effects of hypercorrection for /l/ coarticulation, which more noticeably affects front vowels than back vowels. We thus find there to be a difference in the categorisation and discrimination of front vs back short/long vowel pairs in prelateral contexts which merits further investigation.

These findings contribute to a growing body of research on the coarticulatory effects of /l/ on vowel production in English and support a theory of sound change where reinterpretation of coarticulatory effects leads to language variation [1-5]. Future research on this vowel merger in AusE would benefit from comparing perception and production data, especially in light of Szalay et al.'s [20] findings that perception of durational contrasts in prelateral contexts correlates with production of those contrasts. Future perception experiments should also take into account the confounding effects of minimal pairs in categorisation tasks and neighbourhood effects in the design of nonce words.

6. Method

We wish to thank Geordie Kidd, Katie Carpenter, Felix Kimber, Lizzie Kelly, and the other students in LING40009 for their contributions to the design of the present studies and for contributing to the data collection. Thanks also to Elliot Hillary, Michael Leurs, Benjamin Lynes, and Sarah Zagar for their assistance in recording the experiment stimuli. We also wish to thank Kathryn Hull for their assistance in teaching the first author statistical analysis.

7. References

[1] J. Blevins, 'A theoretical synopsis of evolutionary phonology', *Theoretical Linguistics*, vol. 32, no. 2, pp. 117–165, 2006.

[2] K. Campbell-Kibler, 'Sociolinguistics and perception', *Language and Linguistics Compass*, vol. 4, no. 6, pp. 377-389, 2010.

[3] J. Harrington, F. Kleber, U. Reubold, F. Schiel, and M. Stevens, 'Linking cognitive and social aspects of sound change using agent-based modeling', *Topics in Cognitive Science*, vol. 10, pp. 707–728, 2018, doi: 10.1111/tops.12329.

[4] J. Ohala, 'Sound change is drawn from a pool of synchronic variation', in *Language Change: Contributions to the Study of its Causes*, L. Brevik and E. Jahr, Eds., Berlin: De Gruyter Mouton, pp. 173-198, 1989.

[5] T. Szalay, T. Benders, F. Cox, and M. Proctor, 'Perceptual vowel contrast reduction in Australian English /l/-final rimes', *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 12, no. 1, pp. 1–25, 2021, doi: https://doi.org/10.5334/labphon.185.

[6] F. Cox and S. Palethorpe, 'Australian English', *Journal of the International Phonetic Association*, vol. 37, no. 3, pp. 341–350, 2007.

[7] D. Recasens, 'An EMA study of VCV coarticulatory direction', *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2828–2841, 2002.

[8] T. Szalay, T. Benders, F. Cox, S. Palethorpe, and M. Proctor, 'Spectral contrast reduction in Australian English /l/-final rimes', *Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 1183–1197, 2021, doi: https://doi.org/10.1121/10.0003499.

[9] L. Wade, 'The role of duration in the perception of vowel merger', *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 8, no. 1, pp. 1–34, 2017, doi: https://doi.org/10.5334/labphon.54.

[10] F. Cox, 'Phonetics and Phonology of Australian English', in *Australian English Reimagined: Structure, Features, and Developments*, L. Willoughby and H. Manns, Eds., Abingdon: Taylor and Francis, 2020, pp. 15–33.

[11] U. Altendorf and D. Watt, 'The dialects in the South of England: Phonology', in *Varieties of English*, B. Kortmann and C. Upton, Eds., Berlin: Mouton de Gruyter, 2008, pp. 194–222.

[12] D. Loakes, N. Graetzer, J. Hajek, and J. Fletcher, 'Vowel perception in Victoria: Variability, confusability and listener expectation', in *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, F. Cox, K. Demuth, S. Lin, K. Miles, S. Palethorpe, J. Shaw, and I. Yuen, Eds., Canberra: Australasian Speech Science and Technology Association, 2012.

[13] D. Loakes, J. Clothier, J. Hajek, and J. Fletcher, 'An investigation of the /el/-/æl/ merger in Australian English: A pilot study on production and perception in South-West Victoria"', *Australian Journal of Linguistics*, vol. 34, no. 4, pp. 436–452, 2014.

[14] D. Loakes, J. Clothier, J. Hajek, and J. Fletcher, 'Sociophonetic variation in vowel categorization of australian english', *Language and Speech*, vol. 0, no. 0, p. 00238309231198520, 2023, doi: 10.1177/00238309231198520.

[15] P. Schmidt, C. Diskin-Holdaway, and D. Loakes, 'New insights into /el/-/æl/ merging in Australian English', *Australian Journal of Linguistics*, vol. 41, no. 2, pp. 66–95, 2021.

[16] B. Thomas and J. Hay, 'A pleasant malady: The Ellen/Allan merger in New Zealand English', *Te Reo*, vol. 48, pp. 69–93, 2005.

[17] E. Lewis and D. Loakes, '/ɐlC/-/ɔlC/ Sound change in Australian English: Preliminary res[ɔ]lts', in *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, Canberra: Australasian Speech Science and Technology Association, 2012, pp. 73–76.

[18] H. Oasa, 'Phonology of current Adelaide English', in *Australian English: The Language of a New Society*, P. Collins and D. Blair, Eds., St Lucia: University of Queensland Press, 1989, pp. 271–287.

[19] F. Cox and J. Fletcher, *Australian English Pronunciation and Transcription*. Cambridge: Cambridge University Press, 2017.

[20] T. Szalay, T. Benders, F. Cox, and M. Proctor, 'Production and perception of length contrast in lateral-final rimes', in *Proceedings of the 17th Australasian Conference on Speech Science and Technology*, J. Epps, J. Wolfe, J. Smith, and C. Jones, Eds., Canberra: Australasian Speech Science and Technology Association, 2018, pp. 129–132.

[21] S. Palethorpe and F. Cox, 'Vowel modification in pre-lateral environments', in *Proceedings of the 6th International Seminar on Speech Production*, Canberra: Australasian Speech Science and Technology Association, 2003.

[22] J. Magnuson, B. McMurray, M. Tanenhaus, and R. Aslin, 'Lexical effects on compensation for coarticulation: The ghost of Christmash past', *Cognitive Science*, vol. 27, pp. 285–298, 2003.

[23] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer*. (2024). Accessed: May 02, 2024. [Online]. Available: http://www.praat.org/

[24] M. Davies, 'The Corpus of Contemporary American English (COCA)'.

[25] R. L. Bundgaard-Nielsen, C. T. Best, and M. D. Tyler, 'Vocabulary size is associated with second-language vowel perception performance in adult learners', *Studies in Second Language Acquisition*, vol. 33, no. 3, pp. 433–461, 2011, doi: 10.1017/S0272263111000040.

[26] R. Bundgaard-Nielsen, C. Best, and M. Tyler, 'Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories', *Applied Psycholinguistics*, vol. 32, no. 1, pp. 51–67, 2011, doi: 10.1017/S0142716410000287.

[27] JASP Team, *JASP*. (2024). https://jasp-stats.org/

[28] S. Lin, S. Palethorpe, and F. Cox, 'An ultrasound exploration of Australian English /CVl/ words', in *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, Canberra: Australasian Speech Science and Technology Association, 2012, pp. 105–108.

[29] P. Strycharczuk, D. Derrick, and J. Shaw, 'Locating de-lateralization in the pathway of sound changes affecting coda /l/', *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 11, no. 1, pp. 1–27, 2020, doi: https://doi.org/10.5334/labphon.236.

[30] B. M. Horvath and R. J. Horvarth, 'A multilocality study of a sound change in progress: The case of /l/ vocalization in New Zealand and Australian English', *Language Variation and Change*, vol. 13, no. 1, pp. 37–57, 2001, doi: doi:10.1017/S0954394501131029.

Realisation of Intervocalic /t/ in Australian English: A Snapshot

Thomas Powell-Davies, Rosey Billington

Australian National University

thomas.powell-davies@anu.edu.au, rosey.billington@anu.edu.au

Abstract

In word-medial intervocalic contexts, a common realisation of contemporary Australian English /t/ is a tap, [], but research on past and present realisations of medial /t/ is limited. This study presents a preliminary exploration of the realisation of medial /t/ by 72 speakers in the Mitchell and Delbridge corpus, collected from 1959–1960. Results show a range of variants in use, including taps, fricated and affricate realisations. There are some differences in the distribution of variants according to sex, and in overall release phase durations. The study highlights the need for further phonetic research on variation and change in Australian English consonants.

Index Terms: phonetics, stops, tapping, flapping, VOT, aspiration, frication

1. Background

1.1. Allophones of intervocalic /t/ in English varieties

In varieties of English, the voiceless alveolar plosive /t/ has a range of allophones. Aspirated [t^h] is a canonical variant in word-initial syllable onsets, but word-medially, particularly before unstressed vowels, various other realisations are described. In many varieties of British English, intervocalic glottalised stops and glottal stops [ʔ] are widespread in this context (e.g.[1]), with lenited fricative and affricate realisations in some regions (e.g. [2]). In North American varieties of English, intervocalic alveolar taps[1] [ɾ] are ubiquitous allophones of /t/ (and /d/) word-medially, and can also occur word-finally [4]. Compared to alveolar plosives, alveolar taps are characterised as acoustically short, with closures of around 10–40ms, and typically voiced (e.g. [5] [6]). Intervocalic tapping is also reported in other varieties of English, for example in South Africa, New Zealand, and Australia (e.g. [7]). The majority of studies on intervocalic /t/ in English varieties are based on classifying /t/-realisations as one of several main allophones such as tapped, aspirated, unaspirated, voiced, glottalised, spirantised, and fricated. However, as various authors have noted, categorising phones can be challenging, even with reference to acoustic information; productions of medial /t/ correspond to a wide range of highly gradient realisations, which could be broadly construed as various types of lenition phenomena [8].

1.2. Word-medial intervocalic /t/ in Australian English

Descriptions of contemporary Australian English (AusE) describe tapping of intervocalic /t/ before unstressed vowels as a common process (e.g. [9]). However, it is not yet clear to what extent tapping occurs within different communities or for different individuals, and [10] points out that intervocalic tapping is not reported in the literature on AusE until the 1980s. In sociolinguistic work based on spontaneous speech data collected with adults and teenagers in Sydney from 1977–81, [11] finds intervocalic [ɾ] in use, more often by male speakers of 'Broad' varieties. An affricated [t^s] is also observed, primarily as a word-final pre-pausal allophone, used more often by speakers who are female, of Greek heritage, and who are middle class, and suggested to be a recent development, at the time. Intervocalic tapping is similarly observed in conversational data collected with Brisbane teenagers in the 1980s, at comparable rates for students at a middle class compared to a working class school, with comments that it is not obligatory but that it is also uncommon not to tap in a tapping environment [12]. In a study of /t/ allophones based on conversational and read speech collected in Victoria in the 1990s, [10] reports usage of allophones [t], [t^s], [ʔ], and [ɾ] in various non-initial contexts. For word-medial intervocalic /t/ onsets in unstressed syllables, the majority of realisations are [ɾ], with some use of plosive [t] (the degree of aspiration is not described), and some tokens of [t^s] for the middle but not lower socioeconomic group in the study (where [t^s] here incorporates both fricatives and plosives with fricated release). Other work on /t/ production in Victoria, including intervocalic tokens, also notes occurrences of tapped and fricated realisations and strong sociolinguistic patterning (e.g. [13] [14]), and a recent study of voiceless plosives produced by children aged 5–13 in regional Victoria finds some gendered patterns, for example greater use of fricated variants by young girls, but also comparable rates of intervocalic tapping for boys and girls as the children get older [15]. Spontaneous speech data collected in the 2000s in Perth also finds a range of word-medial intervocalic realisations of /t/, including tapped and fricated variants, with fricated realisations being used more often by speakers from suburbs classified as higher socioeconomic status [16].

In [10] it is noted that tapped and fricated (and glottalised) /t/ allophones are not mentioned in the foundational large-scale survey conducted by Mitchell and Delbridge [M&D] based on data collected from 1959–1960 with nearly 8,000 high school students across Australia [17]. However, [13] note that fricated variants can be observed in the M&D audio-recorded materials [18], and some observations in the M&D written work also suggest other realisations. For example, in some transcribed examples of words such as *water* and *quarter*, intervocalic /t/ is represented as [d̥] and described as 'weakly articulated'.[2] With reference to phonetic transcriptions based on auditory impressionistic analyses, such realisations are used by M&D as minimal indicators of whether speakers show evidence of 'weak' consonant articulations in general, a categorisation which is not

[1] In this paper, we use the term 'tap', but note that 'flap' is also used by some authors for this kind of segment, while others reserve 'flap' for segments such as [ɽ] which require retraction followed by a forward-striking movement of the active articulator (see e.g. [3]).

[2] Specifically, M&D write that "the diacritic [̥] is added ... to indicate a voiceless sound somewhat more weakly articulated than the usual voiceless variety (*lenis* instead of *fortis*)."

given firm parameters but appears to broadly relate to impressions of articulatory strength. This is reportedly the case for 7% of the sample and more common for males and those classified as speakers of the 'Broad' variety of AusE. In 1960s print media, there are also orthographic representations indicative of a non-aspirated or voiced medial /t/ (e.g. 'bewdy' for beauty) [19], prompting questions about the nature of /t/ allophones and the presence of potential tap-like variants in this time period.

1.3. Acoustics of /t/ allophones in Australian English

As [16] note, discrete segmental categorisations of /t/ allophones are abstractions away from the complexities of speech production, and there may be sociolinguistically salient detail to uncover in more fine-grained phonetic analyses. As articulatorily complex sounds, plosives are potentially rich sources of variation, and phonetic analyses for various languages show differences on the basis of factors such as region, gender, and age, though the nature of these differences varies [20]. VOT, which is by far the most frequently analysed measure, shows high variability within and across speakers, but there is also evidence for speaker-specific VOT profiles that hold across e.g. places of articulation and speech rates ([21] [22]). For AusE, relative to the wealth of acoustic phonetic research on vowels, there has been more limited exploration of consonant acoustics. Existing research includes various phonetic examinations of coda /t/, and other coda plosives, in the context of work demonstrating the increasing prevalence of glottalisation in word-final voiceless plosives (e.g. [23]), and analyses of VOT for word-initial /t/ showing influences of factors such as ethnic identity and gender (e.g. [24]). Measures of VOT and closure duration for word-initial and word-medial plosives provide supporting evidence for the phonological contrast between two stop series in mainstream and Indigenous AusE varieties [25], and a detailed acoustic study of fricated /t/ in word-medial and word-final intervocalic contexts shows it is spectrally similar to /ʃ/. While these are not the only studies, there is clear scope for more comprehensive phonetic research on AusE plosives, particularly /t/ in word-medial intervocalic contexts, given the range of realisations possible in this environment.

2. Aim of this study

This study aims to provide an exploratory examination of realisations of word-medial intervocalic /t/ in AusE at the beginning of the 1960s, drawing on the M&D data. The availability of the M&D corpus offers the opportunity to revisit the impressionistic observations in [17] and develop a more detailed snapshot. While the data is limited in scope, the M&D study design has been used as a reference point in the creation of more recent corpora, and a better understanding of medial /t/ characteristics in this early corpus of AusE speech will facilitate future investigation of medial /t/ realisation over time.

3. Method

3.1. Materials

The materials for this study are taken from the corpus collected by Mitchell and Delbridge in 1959–60 [18]. For this preliminary exploration, a subset of the corpus was chosen based on available metadata, which includes school location, birthplace, and speaker sex. The age range of participants in the M&D study was 16–18. The subset included 72 speakers (44 female, 28 male) from three high schools in southern Tasmania, all born in Australia. The M&D corpus includes two sentences that participants were asked to read, designed to inform vowel analyses, but each contains a /t/ context of interest for the present study. The first – *"Let's pick a good spot near the water and spend the morning surfing and relaxing in the sun"* – provides a context with word-medial intervocalic tokens at the onset of an unstressed syllable, taken from 'water'. Three tokens were excluded from this context, as extensive devoicing of following syllables meant their environment was no longer intervocalic. The second sentence – *"The plane flew low over the runway, then increased speed and circled the aerodrome a second time"* – provides a context with word-initial tokens, taken from 'time', as a point of comparison. Altogether, there are 141 tokens – 72 word-initial and 69 word-medial.

3.2. Data processing and annotation

Textgrids were created for each .wav file from the corpus, and /t/ allophones were coded in Praat [26] with reference to waveforms and corresponding spectrograms. The data presented a range of /t/ realisations, which were categorised into variant types with with reference to their acoustic characteristics. These variants and their characteristics had similarities to what has been reported in other works (e.g. [13] [27]). The categorisation was undertaken based on the following criteria:

- **Aspirated [tʰ]** – a voiceless token where the period of closure is followed by a period of h-like frication.
- **Affricate [tˢ]** – a voiceless token where the period of closure is followed by a period of s-like frication.
- **Fricated [t̞]** – a voiceless token where there is no period of closure observed in the waveform, and the frication is observed for the entire duration of the phone.
- **Tapped [ɾ]** – a token with no break in voicing observed on the spectrogram, with the closure duration relatively short.
- **Partially voiced [d̥]** – a token that is voiced for most of the closure. Auditorily, this variant sounds similar to what is often found in English /d/. It is usually followed by a very short release phase. The voicing is what was used to distinguish these from the aspirated tokens. This category has similarities to what has been described as voiced in other work (e.g. [13]).

It is worth noting that the aspirated and affricate variants proved particularly difficult to definitively distinguish. As reported by others who have worked with similar data (e.g. [27]), in some cases there are tokens which show characteristics of both h-like and s-like frication. Additionally, a release burst, as would be typical for [tʰ], was not always clearly identifiable. In these cases, tokens were categorised based on whether they showed more h-like or s-like frication. It is worth emphasising that aspirated [tʰ] and affricate [tˢ] cannot be categorically distinct, as the degree of frication at the alveolar and glottal places is highly gradient. As abstractions from the realities of speech production, these categories are only an initial reference point in delimiting the many gradient possibilities of phonetic detail.

For /t/ realisations where there was an observable occlusion and clear release phase (however short), namely [tʰ], [tˢ], [d̥], the release phase was also annotated, with reference to the onset of a release burst or frication, and the onset of periodicity in the following vowel.

3.3. Analytical procedures

The .Textgrid files and corresponding .wav files were used to create a database in the EMU Speech Database Management System [28], and /t/ variant types and release phase measures for initial and medial tokens were queried using the emuR package in R [29].

4. Results

4.1. Word-initial and intervocalic realisations of /t/

Table 1: *Token numbers for /t/ variant types word-initially (in 'time') and word-medially (in 'water') by sex.*

Phone	Word-initial		Word-medial		**Total**
	Female	Male	Female	Male	
[t^h]	40	20	15	10	**85**
[t^s]	4	2	15	1	**22**
[t̞]	0	0	11	7	**18**
[ɾ]	0	0	0	5	**5**
[d̥]	0	6	0	5	**11**
All	44	28	41	28	**141**

The five categories of /t/ realisation showed different distributions, depending on word position and sex. The token counts are presented in Table 1. In word-initial position, the aspirated realisations were the most common by far (60/72, 80%), with the affricate realisations appearing in only a handful of tokens across both male (2/28, 7%) and female (4/44, 9%) speakers. However, there were six tokens of partially voiced realisations for the male speakers (21%), and none for the female speakers.

In word-medial intervocalic position, there was a similar proportion of aspirated realisations for the female (15/41, 37%) and male (11/28, 39%) speakers. The rest of the tokens for the female speakers comprised affricate (15/41, 37%) and fully fricated (11/41, 27%) realisations. The male speakers had far fewer of these realisations, with seven fricated tokens (25%) and just one affricate token (4%). Instead, they had five each (18%) of both partially voiced and tapped tokens, which were absent for the female speakers.

4.2. Release phase duration by variant type

The duration of the release phase, here referring to the Voice Onset Time of plosive phones as well as the fricated release of affricate phones, is shown for partially voiced [d̥], aspirated [t^h] and affricate [t^s] tokens in Figure 1. The partially voiced tokens have very short release phases word-initially (mean 25ms, σ 6ms, n = 6) and word-medially (mean 26ms, σ 9ms, n = 5), typical of short-lag release (e.g. [30]). The release phases for the aspirated tokens are much longer both word-initially (mean 53ms, σ 17ms, n = 60) and word-medially (mean 45ms, σ 14ms, n = 25). The release phases of the affricate tokens are longer still for both word-initial tokens (mean 57ms, σ 32ms, n = 6) and word-medial tokens (mean 64ms, σ 14ms, n = 16). However, for the affricate tokens, the fricated release phase takes up a greater proportion of the phone relative to the closure phase (initial: mean 64%, σ 16%, n = 6; medial: mean 73%, σ 9%, n = 16) than the Voice Onset Time does in the aspirated tokens (initial: mean 59%, σ 13%, n = 60; medial: mean 58%, σ 13%, n = 25), though recalling that as noted in Section 3.2, these variant types are difficult to strictly delineate.

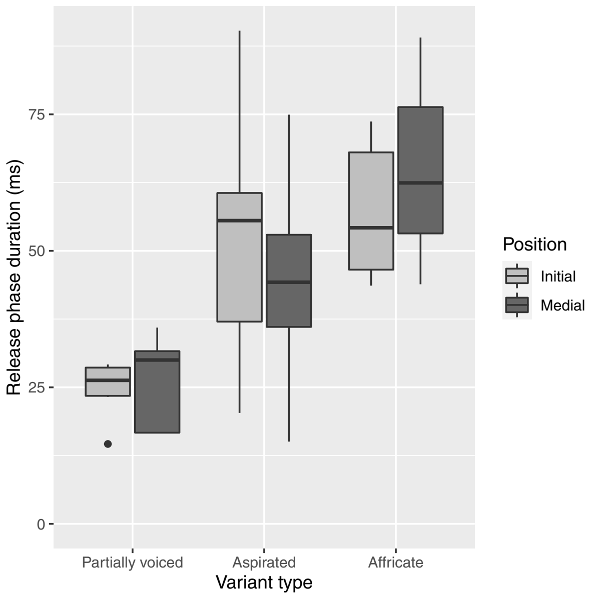


Figure 1: *Release phase duration (ms) for /t/ variants, by variant type ([d̥], [t^h], [t^s]) and word position.*

4.3. Initial vs. medial release phase duration by speaker

The tendency towards longer release phases for initial compared to medial aspirated [t^h] aligns with typical patterns for English voiceless plosives of longer VOT lag for stressed compared to unstressed onsets, and for (word-/utterance-) initial onsets compared to medial onsets (e.g. [31] [32]). Given that there are also (mixed) reports of gender differences in English VOT, and evidence for speaker-specific VOT profiles (Section 1.3), in Figure 2 we examine release phase duration by speaker for initial vs. medial [t^h], [t^s], and [d̥], for all speakers who produced one of these variants in both contexts. There is no connection observed between the release phase in the initial and medial positions (R^2 = 0.05). This is still the case when partially voiced tokens are removed (R^2 = 0.018). However, there are clear differences in the duration of the release phase for female compared to male speakers, with the female speakers tending to have longer release phases.

5. Discussion

This study builds on the auditory analyses of [17] and finds that a range of realisations of /t/ are evident in the M&D sentence data from 1959–1960. This is as expected, given the range of variants observed in more recent studies (e.g. [11] [10]), but not apparent from the higher-level categorisations (e.g. of 'weak' consonants) in the original work, and therefore adds to our understanding of /t/ realisation at an earlier stage of AusE. In particular, the data shows clear evidence that taps, as well as fricated and affricate realisations, were in word-medial use at least two decades before the earliest reports in existing research.

In the word-medial intervocalic position of particular interest here, aspirated realisations are most common, followed by affricate and fricative realisations. Tapped and partially voiced realisations are present in smaller numbers. This is a much more diverse set of variants than found in the word-initial context, in-

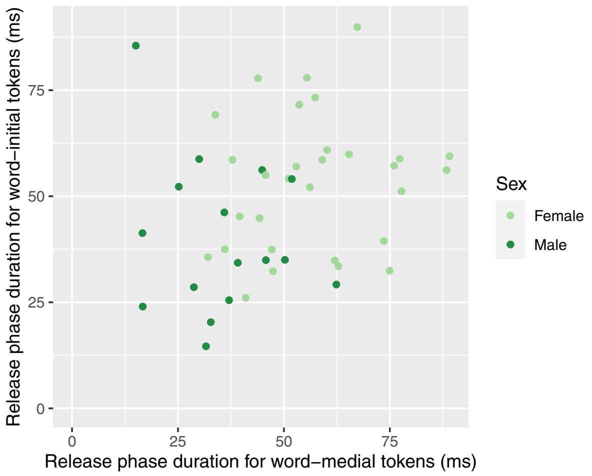


Figure 2: *Release phase duration of [t^h], [t^s], and [d̥] in initial versus medial positions, by speaker and sex.*

dicating higher variability in the word-medial intervocalic context here, when /t/ is the onset of an unstressed syllable. Tapped realisations are much less common than more recent reports suggest (e.g. [12]), noting however that the speech style of the two elicited sentences is likely to differ from the more conversational style of speech analysed in some later studies.

The use of different variants in this data appears to differ according to speaker sex. Both female and male speakers produced aspirated, affricate and fricated tokens, but none of the female speakers produced any of the partially voiced or tapped variants, whereas these were well represented in the data for the male speakers. This has similarities to findings reported by [11] on the greater use of taps by 'Broad' male speakers.

The differences in variant distribution may be related to the differences in release phase duration. While there appears to be no correlation between a speaker's release phase duration in word-initial position and word-medial intervocalic position (as might have been expected given other kinds of evidence for speaker-specific VOT profiles [22]), it is apparent that the male speakers have shorter release phases in general. This has also been observed, albeit inconsistently, in other varieties of English [20]. The shorter release phases and the presence of partially voiced tokens for male speakers could potentially show the earlier phases of a process that has progressed into the widespread distribution of taps observed in present-day Australian English [9]. However, much more detailed and comprehensive analyses, together with direct comparison with data from later timepoints, will be needed to explore this.

As part of more detailed and comprehensive analyses, further examination of the phonetic characteristics of word-medial intervocalic /t/ will be of value, and complement the increasing number of phonetic studies on word-initial and word-final /t/. The precise processes at play in this environment, at the onset of unstressed syllables, can be difficult to characterise, given that the phone realisations are highly gradient (and arguably correspond to a range of different types of lenition, e.g. [10] [8]). However, additional fine-grained phonetic detail is likely to significantly add to our understanding of socially and linguistically structured variation for AusE /t/ and other plosives.

6. Conclusion

The data presented sheds light on the realisation of /t/ in Australian English, with a focus on word-medial intervocalic segments at the onset of unstressed syllables. It shows that multiple variants of intervocalic /t/ were already in use at the time Mitchell & Delbridge collected data for their foundational study of Australian English speech, well before this range of variants was described in the scholarly literature. This raises questions about the nature of /t/ realisation in Australian English over time, and to what extent there is evidence for stable variation compared to changing patterns in /t/ realisation. More broadly, this study also highlights the need to continue adding to sociophonetic research on consonants in Australian English, to complement the extensive research on variation and change in Australian English vowels.

7. Acknowledgements

We wish to thank two anonymous reviewers for their generous comments.

8. References

[1] Docherty, G. J. and Foulkes, P. "Glottal variants of /t/ in the Tyne-side variety of English", in W. J. Hardcastle, and J. Beck, [Eds], *A figure of speech: A festschrift for John Laver*, 213–240. Routledge, 2014.

[2] Marotta, G., Barth, M., et al. "Acoustic and sociolinguistic aspects of lenition in Liverpool English", Studi Linguistici e Filologici On Line, 3(2):377–413, 2005.

[3] Ladefoged, P. and Maddieson, I. The sounds of the world's languages. Blackwell, 1996.

[4] De Jong, K. J. "Flapping in American English", in Oostendorp, M., Ewen, C. J., Hume, E., and Rice, K., editors, *The Blackwell Companion to Phonology*, 1–19. Wiley, 1 edition, 2011.

[5] Zue, V. W. and Laferriere, M. "Acoustic study of medial /t, d/ in American English", Journal of the Acoustical Society of America, 66(4):1039–1050, 1979.

[6] Turk, A. "The American English flapping rule and the effect of stress on stop consonant durations", Cornell Working Papers in Phonetics, 7:103–133, 1992.

[7] Wells, J. C. Accents of English, volume 1–3. Cambridge University Press, 1982.

[8] Ashby, M. and Przedlacka, J. "The stops that aren't", Journal of the English Phonetic Society of Japan, 1:14–15, 2011.

[9] Cox, F. and Fletcher, J. Australian English pronunciation and transcription. Cambridge University Press, 2 edition, 2017.

[10] Tollfree, L. "Variation and change in Australian English consonants: Reduction of /t/", English in Australia, 26:45–67, 2001.

[11] Horvath, B. M. Variation in Australian English: The sociolects of Sydney. Cambridge University Press, 1985.

[12] Ingram, J. "Connected speech processes in Australian English", Australian Journal of Linguistics, 9(1):21–49, 1989.

[13] Loakes, D., McDougall, K., and Gregory, A. "Variation in /t/ in Aboriginal and Mainstream Australian Englishes", in *Pro-ceedings of the Eighteenth Australasian International Conference on Speech Science and Technology*, 61–65. Australasian Speech Science and Technology Association, 2022.

[14] Loakes, D. and McDougall, K. "Individual variation in the frication of voiceless plosives in Australian English: A study of twins' speech", Australian Journal of Linguistics, 30(2):155–181, 2010.

[15] Ford, C. Acquisition of gender-specific sociophonetic cues in the speech of primary school-aged children. PhD Thesis, La Trobe University, 2018.

[16] Docherty, G., Foulkes, P., Gonzalez, S., and Mitchell, N. “Missed connections at the junction of sociolinguistics and speech processing”, Topics in Cognitive Science, 10(4):759–774, 2018.

[17] Mitchell, A. G. and Delbridge, A. The speech of Australian adolescents: A survey. Angus & Robertson, 1965.

[18] Mitchell, A. G. and Delbridge, A. The speech of Australian adolescents: Research data and recordings collected by A.G. Mitchell and Arthur Delbridge in 1959 and 1960 [available via https://hdl.handle.net/2123/31585], 1998. University of Sydney.

[19] Army: The Soldier’s newspaper. Here’s a ‘Bewdy’, Nov 1964.

[20] Chodroff, E. and Foulkes, P. “Sociophonetics and stops”. In Strelluf, C., editor, *The Routledge handbook of sociophonetics*, pages 143–175. Routledge, 2024.

[21] Theodore, R. M., Miller, J. L., and DeSteno, D. “Individual talker differences in voice-onset-time: Contextual influences”, The Journal of the Acoustical Society of America, 125(6):3974–3982, 2009.

[22] Chodroff, E. and Wilson, C. “Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English”, Journal of Phonetics, 61:30–47, 2017.

[23] Penney, J., Cox, F., Miles, K., and Palethorpe, S. “Glottalisation as a cue to coda consonant voicing in Australian English”, Journal of Phonetics, 66:161–184, 2018.

[24] Clothier, J. and Loakes, D. “Coronal stop VOT in Australian English: Lebanese Australians and mainstream Australian English speakers”. In *Proceedings of the Seventeenth Australasian International Conference on Speech Science and Technology*, pages 13–16. Australasian Speech Science and Technology Association, 2018.

[25] Mailhammer, R., Sherwood, S., and Stoakes, H. “The inconspicuous substratum: Indigenous Australian languages and the phonetics of stop contrasts in English on Croker Island”, English World-Wide, 41(2):162–192, 2020.

[26] Boersma, P. and Weenink, D. Praat: doing phonetics by computer, 2022.

[27] Buizza, E. and Plug, L. “Lenition, fortition and the status of plosive affrication: the case of spontaneous RP English”, Phonology, 29(1):1–38, 2012.

[28] Winkelmann, R., Harrington, J., and Jänsch, K. “EMU-SDMS: Advanced speech database management and analysis in R”, Computer Speech & Language, 45:392–410, 2017.

[29] R Core Team. R: A Language and Environment for Statistical Computing, 2021.

[30] Cho, T. and Ladefoged, P. “Variation and universals in VOT: Evidence from 18 languages”, Journal of Phonetics, 27(2):207–229, 1999.

[31] Lisker, L. and Abramson, A. S. “Some effects of context on voice onset time in English stops”, Language and Speech, 10(1):1–28, 1967.

[32] Cho, T. and Keating, P. “Effects of initial position versus prominence in English”, Journal of Phonetics, 37(4):466–485, 2009.

Voice Quality in Voiceless Coda Stop Contexts: Evidence from Australian English Speakers with English and Arabic Language Backgrounds

Joshua Penney, Felicity Cox

Department of Linguistics, Macquarie University, Australia

joshua.penney@mq.edu.au felicity.cox@mq.edu.au

Abstract

Voice quality has the potential to index social characteristics but it remains understudied in sociophonetics. We examined voice quality in vowels preceding voiceless coda stops in Australian English speakers from monolingual English backgrounds and Arabic language backgrounds. In general, the Arabic background group used breathier voice quality (higher spectral tilt). We also found voice quality differences related to coda place of articulation. While both groups showed evidence of glottal constriction preceding /t/ and breathiness preceding /k/, the Arabic background group produced increased breathiness in non-alveolar contexts compared to the English background group, suggesting laryngeal setting differences for achieving voicelessness.

Index Terms: voice quality, phonation, voiceless stops, Australian English, ethnolects.

1. Introduction

In recent years, there has been increasing attention in research on Australian English (AusE) to how phonetic variation may be linked to speakers' ethnolinguistic backgrounds. A particular focus has been on how the use of certain phonetic features may index a 'non-mainstream' identity, or a level of divergence from speakers of the mainstream variety of AusE (MAusE) that is spoken by the majority of the population [1, 2, 3, 4]. Although not the only language background group to be included in such analyses, speakers with Arabic language backgrounds (particularly Lebanese Arabic) have been shown to exhibit differences to MAusE speakers for a range of features: for example, the timing of constituent components of VC rhymes [5], voice onset time [6], degree of /l/ velarisation [7], and the use of glottalisation as a hiatus breaking strategy [8]. A complete account of the features involved and whether the identified differences from MAusE represent a specific ethnolect, or rather form part of a more general multiethnolectal variety of AusE, is not yet well understood and is the subject of ongoing analysis [e.g. 9].

Previous research has provided some indication that in addition to the features described above, differences in voice quality may also exist between AusE speakers with Arabic language backgrounds and MAusE speakers. In an analysis of monosyllabic CV words, [10] found an overall breathier voice quality in speakers with Arabic backgrounds compared to MAusE speakers. [10] also identified that Arabic background speakers produced breathier offsets at the right edge of CV syllables, which they speculated may be due to phonotactic influences from Arabic.

Voice quality differences have been identified more broadly between speakers of mainstream varieties and those with non-mainstream backgrounds in a number of language varieties. In New Zealand English, [11] found that speakers with Māori backgrounds had both higher F0 and more creaky voice compared to Pakeha (i.e. white European) speakers (cf. [12], who also found a voice quality difference between Māori and Pakeha speakers but in the opposite direction). [13] identified a breathier voice quality as a feature of American English speakers with Chinese and Korean backgrounds that contributes to them "sounding Asian". In British English, [14] found that male speakers of the multiethnolectal variety Multicultural London English produced speech with a breathier voice quality than Anglo speakers from outside of London, and in a subsequent analysis showed that creaky voice was a feature of Anglo speakers from outside London [15]. A breathier voice quality has also been identified in speakers of the multiethnolectal German variety Kiezdeutsch compared to speakers of Standard German [16]. With regard to AusE, [17] found that male speakers of Aboriginal Australian English produce lower F0 and a creakier voice quality compared to MAusE speakers. In an analysis specifically examining creaky voice prevalence in different areas of Sydney, [18] found that the use of creaky voice varied according to gender and area, pointing towards potential differences within and across the dominant ethnic groups who live in these different areas.

Differences in voice quality are also known to result from phonological context in AusE: glottalisation is reported to occur in association with voiceless coda stops, resulting in laryngealisation (i.e. a period of creaky voice) of preceding vowels [19, 20]. While this has primarily been studied in coda /t/ contexts, [21] found acoustic evidence for glottalisation preceding voiceless coda stops occurring at all three places of articulation (/p, t, k/) in the unstressed syllables of trochees. [22] also found evidence of laryngealisation in voiceless coda stop contexts at all three places of articulation in the speech of AusE children. However, more recent analyses utilising electroglottography (EGG) found that the voice quality of vowels preceding voiceless coda stops in monosyllabic words varied according to stop place of articulation [23, 24]: vowels preceding /t/ and (to a lesser extent) /p/ showed increased glottal constriction towards the offset of the vowel, whereas increased glottal spreading was observed in vowels preceding /k/.

The results from the EGG analyses [23, 24] suggest that different strategies may be used to achieve voicelessness depending on the place of articulation of the coda stop, and this may result in different voice qualities during phonation for the preceding vowel: breathiness (or increased glottal spreading) preceding /k/ and creakiness (or increased glottal constriction) preceding /t, p/. However, these findings were based on speakers of MAusE with monolingual English backgrounds only. It remains an open question as to whether speakers with non-mainstream AusE backgrounds also exhibit this differential pattern of achieving voicelessness. To address this question, in this study we analyse voice quality in vowels occurring in voiceless coda stop contexts in two groups of speakers: AusE speakers with monolingual English backgrounds and AusE speakers with Arabic language backgrounds.

2. Methods

2.1. Participants and data collection

The data for this analysis were extracted from the Multicultural Australian English –Voices of Sydney (MAE-VoiS) project [9], in which the speech of 183 adolescents from various areas of Sydney was recorded. The participants in MAE-VoiS were sampled from parts of Sydney that differ in the level of linguistic diversity in the community and the dominant non-English languages spoken. Participants' speech was recorded as they engaged in a picture naming elicitation task and a spontaneous conversation with a peer facilitated by a research assistant from the community. Full details of the corpus, participants, and recording are available in [25].

For this analysis, a subset of 23 male participants aged 15-16 years (mean = 15.6 years) were selected to enable a comparison of speakers from monolingual English backgrounds and from Arabic language backgrounds. Only male speakers are examined as the corpus contains only a small number of female speakers with Arabic language background. 10 of the speakers were from monolingual English backgrounds; the other 13 were from Arabic language backgrounds. Note that the speakers with Arabic backgrounds varied in terms of their proficiency in and use of Arabic. All of the speakers included in this analysis were recorded in a face-to-face setting via a Zoom H6 digital recorder through a Rode HS2 headset microphone, with 44.1kHz sampling rate and 16-bit quantisation.

Only data collected in the picture naming task is included in this analysis. For each speaker, we extracted their productions of a set of single words with coda voiceless stops /p, t, k/. Each voiceless coda was preceded by a monophthongal vowel, with all of the AusE monophthongs sampled except for /eː/, which occurs in open monosyllables or followed by alveolar consonants only [26]. Onsets varied across the words. Each word was produced once by each participant; however, in some cases participants did not produce all of the words and in other cases a single word may have been repeated. All repetitions were included. In total, 615 items are included in the analysis. Table 1 provides an overview of the number of items according to coda context and language background group. We note that the bilabial coda context is underrepresented compared to alveolar and velar, as will be discussed below.

Table 1. Number of items according to voiceless coda context and language background group.

Group	Bilabial	Alveolar	Velar	Total
Arabic	39	205	103	347
English	30	158	80	268
Total	**69**	**363**	**183**	**615**

2.2. Acoustic analysis

All items were processed through WebMaus [27] to provide segmentation and forced alignment at the level of the phoneme. Segment boundaries were subsequently checked and hand corrected by trained research assistants. Acoustic measures of voice quality were extracted using VoiceSauce [28]. H1*-H2* is a measure of spectral tilt that is correlated with degree of glottal constriction: higher values of H1*-H2* indicate increased glottal opening whereas lower values of H1*-H2* indicate increased glottal constriction [29, 30, 31, 32]. H1*-H2* was estimated in 1 millisecond increments across each vowel. F0 measures were estimated using REAPER [33] through a custom script implemented in VoiceSauce [28]. Accurate estimation of harmonics requires reliable F0 estimation; REAPER produces robust F0 estimation even in the presence of irregular voicing and low pitch that may occur during creaky phonation [33]. The presence of nearby vowel formants can increase the amplitude of the harmonics. Therefore, in order to enable comparison across different vowel qualities, a correction algorithm was applied [34], as indicated by the asterisks (H1*-H2*). Formant measures for the amplitude correction were estimated by Praat [35] with default settings. All values were time normalised across the duration of the vowel.

2.3. Statistical analysis

The data were modelled with a generalised additive mixed model (GAMM), implemented in the mgcv [36] and itsadug [37] packages in R [38]. To examine the interaction between language background and place of articulation (POA), we created a hardcoded interaction variable: Lang-POA, with the following levels: English-Alveolar, English-Bilabial, English-Velar, Arabic-Alveolar, Arabic-Bilabial, Arabic-Velar. This was an ordered factor with English-Alveolar as the reference level. Lang-POA was included as a parametric effect in the model, to examine overall effects language background and POA. In order to examine potential differences in H1*-H2* trajectory shape, the model also included a reference smooth over normalised time, and a smooth over normalised time by Lang-POA, fitted with thin plate regression splines with 10 knots. We also included a random effect of word to account for the different onsets and different vowels present in the items, and a factor smooth over normalised time by participant, to account for individual speaker differences.

(1)

bam(T1 ~ Lang-POA + s(normalised_time) +
s(normalised_time, by= Lang-POA, bs="tp" k = 10) +
s(word, bs = "re") + s(normalised_time, participant, bs =
"fs", m = 1)

An AR1 error model was incorporated into the final model to account for the fact that autocorrelation is likely to be present in vowel trajectory data [39, 40]. The model code is shown in (1) below. To reduce the likelihood of increased type I error, we follow [39] and consider effects significant at an alpha of $p <$ 0.025 in our interpretation of the model results.

3. Results

Table 2 displays mean H1*-H2* values for each language background according to POA and overall. As can be seen, the Arabic background group exhibited higher mean H1*-H2* values compared to the English background group both overall and in each POA, indicative of less glottal constriction in this group.

Table of Contents
for this manuscript

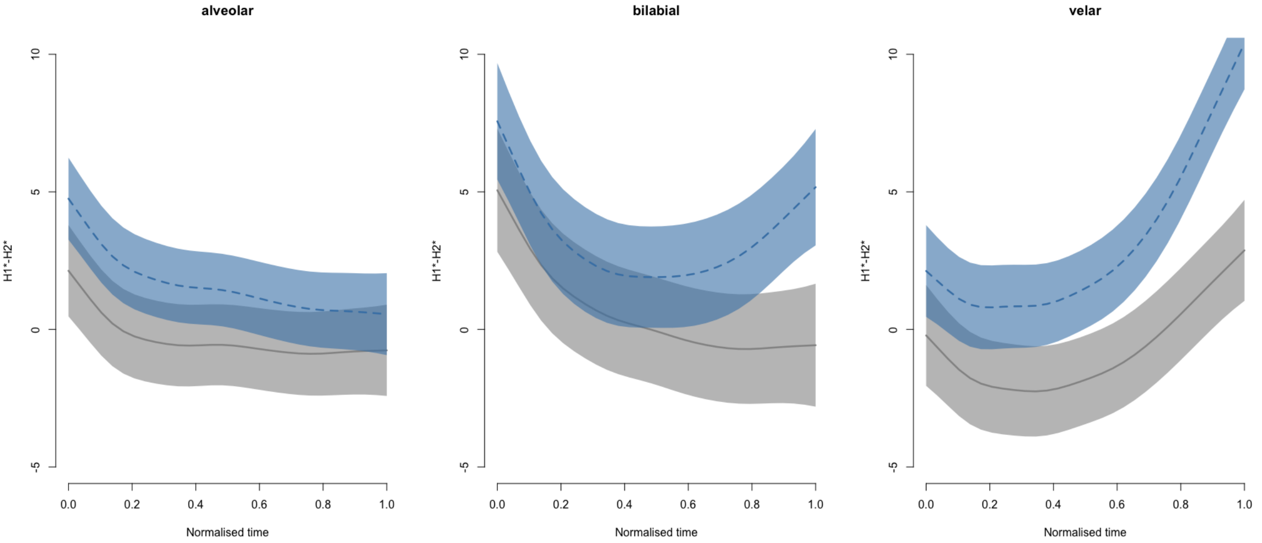


Figure 1. Model predictions of H1-H2* across the vowel preceding alveolar (left panel), bilabial (middle panel), and velar (right panel) coda contexts for speakers with monolingual English (solid grey lines) and Arabic (dashed blue lines) language backgrounds. Error ribbons show 95% CIs.*

Table 2. Mean H1-H2* values (dB) according to place of articulation and language background group.*

Group	Bilabial	Alveolar	Velar	Overall
Arabic	2.8	1.58	2.97	2.13
English	0.2	-0.46	-0.65	-0.46

The results of the GAMM are given in Table 3. Compared to the reference of English-Alveolar, there was no significant parametric difference in either /p/ or /k/ for the English background group. For the Arabic background group, there was no significant parametric difference from the reference for the alveolar POA (at an alpha of p < .025), but there were significant parametric differences for both bilabial and velar POAs. This indicates that overall differences between POAs were not significant within the English background group, nor were overall differences significant between the groups for the alveolar POA (although this might be considered a trend at p = .033). On the other hand, the Arabic group had overall significantly higher H1*-H2* in the bilabial and velar POAs compared to the reference. Pairwise comparisons using [41] show that the differences between the groups were significant for the velar (p = 0.003) but not the bilabial POA (p = 0.08).

Turning to the smooth terms, within the English background group the smooths for both bilabial and velar POAs were significantly different from the reference. For the Arabic background group, the smooth for the alveolar POA did not differ significantly from the reference, but those for the bilabial and velar POAs did. This indicates that the H1*-H2* trajectories differ across the three POAs in both of the groups, although there are differences between the groups in how these differences manifest. This can be seen in Figure 1: in the left panel, which represents the alveolar context, both groups exhibit a similar decrease in H1*-H2* from early in the vowel, with H1*-H2* remaining low throughout the rest of the trajectory. In the velar context, shown in the right panel, both groups show an initial drop in H1*-H2* followed by an increase from roughly midway through the vowel's trajectory. However, the Arabic background group demonstrate a sharper increase towards the end of the vowel's trajectory than the English background group. The most striking difference between the groups is evident in the bilabial context, shown in the middle panel: the English background group show a gradual decrease in H1*-H2* across the trajectory of the vowel, beginning from a slightly higher position compared to the alveolar context; the Arabic background group also show a similar decrease at the start of the vowel, but this diverges from the English background group's trajectory with an increase from approximately the middle of the vowel, similar to in the velar context. Inspection of the estimated differences between smooths for the two groups confirm that the differences are significant (i.e. confidence intervals show no overlap with zero [39]) throughout the vowel in the velar context, and in the second half of the vowel in the bilabial context.

4. Discussion

The results described above accord with recent findings from articulatory analysis using EGG; namely, that MAusE speakers appear to be use different strategies to achieve coda stop voicelessness depending on the place of articulation of the coda stop, and that these strategies have implications for voice quality produced during the preceding vowel. As reported in [23, 24], we here observed acoustic evidence of increased glottal constriction in vowels preceding coda /t/ and coda /p/ for speakers in the English language background group. On the contrary, acoustic evidence of increased breathiness was observed for these speakers in coda /k/ contexts, suggesting voicelessness in this context is achieved through glottal spreading [23]. While glottal constriction in coda /t/ contexts, and to a lesser extent /p/ contexts, is consistent with previous reports of glottalisation in AusE [19, 20], a few previous studies have also found acoustic evidence of glottal constriction occurring in coda /k/ [21, 22], which differs from the pattern observed here. While we note that the focus of [21, 22] differed somewhat from this analysis and the EGG study reported above—one study analysed the unstressed syllables of trochaic feet and the other focused on children's speech—it remains an empirical question as to why conflicting results have been observed. We leave this to future research to investigate.

Table 3. Summary of the results of the generalised additive mixed model.

Parametric coefficients	Estimate	SE	*t*	*p*
(Intercept)	-0.446	0.739	-0.603	0.546
English-Bilabial	0.874	0.707	1.236	0.217
English-Velar	-0.482	0.485	-0.993	0.321
Arabic-Alveolar	1.969	0.926	2.127	0.033
Arabic-Bilabial	3.474	1.146	3.032	0.002
Arabic-Velar	3.338	1.026	3.252	0.001
Smooth terms	edf	Ref. df	*F*	*p*
s(normalised time)	6.238	7.218	6.918	<0.0001
s(normalised time): English-Bilabial	2.024	2.549	4.545	0.006
s(normalised time): English-Velar	3.430	4.389	31.783	<0.0001
s(normalised time): Arabic-Alveolar	1.001	1.001	2.110	0.146
s(normalised time): Arabic-Bilabial	3.289	4.232	6.026	<0.0001
s(normalised time):Arabic-Velar	4.741	5.978	34.114	<0.0001
s(word)	21.362	24.0	9.506	<0.0001
s(normalised time, participant)	100.286	205.0	5.880	<0.0001

While the pattern of increased glottal constriction in /t/ and /p/ contexts and increased breathiness in /k/ contexts in English language background speakers corresponds with previous articulatory research on MAusE, we observed divergence from this pattern in the speakers with Arabic language backgrounds. These speakers exhibited a pattern of glottal constriction in the /t/ context but breathiness in both the /p/ and /k/ contexts, indicating group level differences in how coda voicelessness is achieved in specific phonological contexts. According to the results shown in this analysis, speakers with English language backgrounds and speakers with Arabic language backgrounds both realise coda voicelessness through increased glottal constriction (resulting in laryngealisation/creakiness on the preceding vowel) in alveolar contexts and, conversely, through increased glottal spreading (resulting in breathiness on the preceding vowel) in velar contexts. In bilabial contexts, speakers with English language backgrounds use glottal constriction, whereas speakers with Arabic language backgrounds use glottal spreading. It should here be pointed out that the bilabial context was not as well sampled as either of the other two places of articulation, as shown in Table 1. Therefore, the interpretation of the results specific to this context should be treated with caution, particularly as this is the very context where the largest differences between the groups were observed.

Nevertheless, the Arabic language background group showed evidence of greater breathiness compared to the English language background group in the velar context, and this group also exhibited higher H1*-H2* values overall in each place of articulation (although with an alpha of p = .025 this difference was not significant in the alveolar context, there was a trend towards significance). These results are consistent with previous research that found evidence for increased breathiness in the vowels and at the offset of CV syllables in Arabic background speakers. We therefore tentatively suggest that breathiness may be a more general voice quality feature indicative of this group [10].

We acknowledge that this analysis is based on a relatively small sample of only male adolescent speakers producing speech in a highly constrained task. It remains to be seen if differences in voice quality such as those observed here will also be present in a larger, more varied, sample of speakers, including female speakers, different age groups, and in spontaneous speech. Future research should also examine whether differences in voice quality are due to influences from a speaker's heritage language, whether these differences might be due to social or rather articulatory differences, and to what extent differences in voice quality are encoded in listeners' perception of different ethnolinguistic identities. We note that a breathier voice quality has been identified in some multiethnolectal varieties, for example speakers of Multicultural London English [14] and Kiezdeutsch [16], but it has also been linked to a specific 'Asian' ethnolect in American English speakers [13]. As noted above, differences in voice quality (specifically creaky voice) have previously been identified in speakers from different areas of Sydney [18]. It will therefore be interesting to examine whether increased breathiness is evident in AusE speakers with other non-English language backgrounds, or if this is specific to those with Arabic language backgrounds.

5. Conclusion

In this study, we have shown voice quality differences in vowels preceding voiceless coda stops, indicating differential strategies for achieving coda voicelessness in AusE speakers according to the place of articulation of the coda stop. Additionally, we observed differences according to speakers' language background: speakers with Arabic backgrounds have an overall breathier voice quality and exhibit a different strategy in bilabial coda contexts compared to those with monolingual English backgrounds. Future research will examine the extent to which such differences in voice quality contribute to a (multi)ethnolectal variety of AusE.

6. Acknowledgements

This research was supported by an Australian Research Council grant (FT180100462) to the second author.

7. References

[1] Clothier, J., "Ethnolectal variability in Australian Englishes", in L. Willoughby and H. Manns [Eds.], Australian English reimagined: Structure, features and developments, 155-172, Routledge, 2019.

[2] Clyne, M., Eisikovits, E. and Tollfree, L., "Ethnic varieties of Australian English", in D. Blair and P. Collins [Eds.], English in Australia, 223-238, Benjamins, 2001.

[3] Cox, F. M., "Australian English: Phonetics and Phonology", in L. Willoughby and H. Manns [Eds.], Australian English Reimagined: Structure, Features and Developments, Taylor & Francis/Routledge, 2020.

[4] Grama, J., Travis, C. and Gonzalez, S., "Ethnolectal and community change ov(er) time: Word-final (er) in Australian English", Australian Journal of Linguistics, 40(3):346-368, 2020.

[5] Cox, F. and Palethorpe, S., "Timing differences in the VC rhyme of standard Australian English and Lebanese Australian English", Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, 528-531, 2011.

[6] Clothier, J. and D. Loakes, “Coronal stop VOT in Australian English: Lebanese Australians and mainstream Australian English speakers”, Proceedings of 17th Australasian International Conference on Speech Science and Technology, Sydney, Australia, 13-16, 2018.

[7] Clothier, J., “A sociophonetic analysis of /l/ darkness and Lebanese Australian Ethnic Identity in Australian English”, Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia, 1888-1892, 2019.

[8] Cox, F., Penney, J. and Palethorpe, S., “Fifty years of change to prevocalic definite article allomorphy in Australian English”, Journal of the International Phonetic Association, 53(3):804-834, 2023.

[9] Cox, F. and Penney, J., “Multicultural Australian English: Voices of Sydney (MAE-VoiS)”, 2024. Available: https://doi.org/10.25949/25572153.v1.

[10] Penney, J. and Cox, F., “Vowel and voice quality differences between mainstream and non-mainstream Australian English speakers”. Paper presented at Forum on Englishes in Australia, La Trobe University, 2021.

[11] Szakay, A., “Voice quality as a marker of ethnicity in New Zealand: From acoustics to perception. Journal of Sociolinguistics”, 16(3):382-397, 2012.

[12] Szakay, A. and King, J., “Voice quality transfer effects between English and Māori”. Paper presented at the 22nd Sociolinguistics Symposium, Auckland, New Zealand, 2018.

[13] Newman, M. and Wu, A., “Do you sound Asian when you speak English? Racial identification and voice in Chinese and Korean Americans’ English”, American Speech, 86:152-178, 2011.

[14] Szakay, A. and Torgersen, E. N., “An acoustic analysis of voice quality in London English: The effect of gender, ethnicity and f0”, Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, Scotland, 2015.

[15] Szakay, A. and Torgersen, E., “A re-analysis of F0 in ethnic varieties of London English using REAPER”, Proceedings of the International Congress of Phonetic Sciences, Melbourne, Australia, 1675-1878, 2019.

[16] Penney, J., Weirich, M. and Jannedy, S., “Increased breathiness in adolescent Kiezdeutsch speakers: A marker of mulitethnolectal group affiliation?”, Language and Speech, in press.

[17] Loakes, D. and Gregory, A., “Voice quality in Australian English”, *JASA Express Letters*, 2(8), 2022.

[18] White H., Penney, J., Gibson, A., Szakay, A. and Cox, F., “Creaky voice prevalence across Sydney”, Journal of the Acoustical Society of America, 154:A335, 2023.

[19] Penney, J., Cox, F., Palethorpe, S. and Miles, K. “Glottalisation as a cue to coda stop voicing”, Journal of Phonetics, 66:161-184, 2018.

[20] Penney, J., Cox, F. and Szakay, A., “Glottalisation, coda voicing and phrase position in Australian English”, Journal of the Acoustical Society of America, 148:3232-3245, 2020.

[21] Penney, J., Cox, F. and Szakay, A., “Glottalisation of word-final stops in Australian English unstressed syllables”, Journal of the International Phonetic Association, 51:229-260, 2021.

[22] Millasseau, J., Bruggeman, L., Yuen, I. and Demuth, K., “Temporal cues to onset voicing contrasts in Australian English-speaking children”, Journal of Child Language, 48(6): 1262-1280, 2021.

[23] Ratko, L., Penney, J. and Cox, F., “Opening or closing? An electroglottographic analysis of voiceless coda consonants in Australian English”, Proceedings of INTERSPEECH 2023, Dublin, Ireland, 1823-1827, 2023.

[24] Ratko, L., Penney, J. and Cox, F., “An electroglottographic study on the effect of following context on glottal constriction in Australian English coda /t/”, Journal of the Acoustical Society of America, 154:A244, 2023.

[25] Cox, F. and Penney, J., “Multicultural Australian English: The new voice of Sydney”, Australian Journal of Linguistics, 2024.

[26] Harrington, J., Cox, F. and Evans, Z., “An acoustic phonetic study of broad, general, and cultivated Australian English vowels”, Australian Journal of Linguistics, 17:155-184, 1997.

[27] Kisler, T., Reichel, U. and Schiel, F., “Multilingual processing of speech via web services”, Computer Speech and Language, 45:326-347, 2017.

[28] Shue, Y.-L., Keating, P., Vicenik, C. and Yu, K. “VoiceSauce: A program for voice analysis,” Proceedings of the International Congress of Phonetic Sciences, Hong Kong, 1846–1849, 2011.

[29] Hillenbrand, J., Cleveland, R. and Erickson, R., “Acoustic correlates of breathy vocal quality”, Journal of Speech, Language, and Hearing Research, 37:769-778, 1994.

[30] Holmberg, E. B., Hillman, R. E., Perkell, J., Guiod, P. and Goldman, S. L., “Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice”, Journal of Speech, Language, and Hearing Research, 38: 1212-1223, 1995.

[31] Garellek, M., “The phonetics of voice”, in W. F. Katz and P. F. Assmann [Eds.], The Routledge Handbook of Phonetics, 75-106, Routledge, 2019.

[32] Keating, P., Kuang, J., Garellek, M., Esposito, C. M. and Khan, S. D., “A cross-language acoustic space for vocalic phonation distinctions”, Language, 99(2):351-389, 2023.

[33] Talkin, D., “REAPER: Robust epoch and pitch EstimatoR,” 2015, [Computer program]. Available: https://github.com/google/REAPER.

[34] Iseli, M., Shue, Y.-L. and Alwan, A., “Age, sex, and vowel dependencies of acoustic measures related to the voice source”, Journal of the Acoustical Society of America, 121:2283-2295, 2007.

[35] Boersma, P., and Weenink, D., “Praat: Doing phonetics by computer” (6.4.04), http://www.praat.org, 2024.

[36] Wood, S N., “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models”, Journal of the Royal Statistical Society, 73:3-36, 2011.

[37] van Rij, J., Wieling, M., Baayen, H. and van Rijn, H., 2020, “itsadug: Interpreting Time series and autocorrelated data using GAMMs”, version 2.4, 2020 [R package]. Available: https://cran.r-project.org/web/packages/itsadug.

[38] R Core Team, “R: A language and environment for statistical computing”, version 4.3.3, 2024 [Computer program]. Available: https://www.R-project.org/.

[39] Soskuthy, M., “Evaluating generalised additive mixed modelling strategies for dynamic speech analysis”, Journal of Phonetics, 84, 101017, 2021.

[40] Wieling, M., “Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English”, Journal of Phonetics, 70:86-116, 2018.

[41] Lenth, R., “emmeans: Estimated marginal means, aka least-squares means”, version 1.4.8, 2020 [R package]. Available: https://CRAN.R-project.org/package=emmeans.

Patterns of Silent Pausing in Aboriginal and Mainstream Australian Englishes Spoken in Warrnambool

Kirsty McDougall[1], *Alice Paver*[1], *Martin Duckworth*[2], *Liz Blackwell*[1] *and Debbie Loakes*[3]

[1]University of Cambridge, [2]Independent Researcher, [3]The University of Melbourne

[1]kem37|aegp2|ecb76@cam.ac.uk; [3]dloakes@unimelb.edu.au

Abstract

A study of variation in the production of silent pauses, both turn-internal and response latencies (silent pauses immediately following the interlocutor's turn) in Aboriginal and Mainstream Australian Englishes (AAE and MAE) spoken in Warrnambool is presented. Neither frequency nor duration of turn-internal silent pauses yielded differences between the varieties. However, response latencies were more frequent in AAE. Further, when the number of opportunities for a response latency to occur were examined, delayed responses appeared more frequently in AAE than MAE. Implications for courtroom discourse contexts are discussed.

Index Terms: silent pauses, response latencies, fluency, disfluency, TOFFA, Australian Englishes, legal discourse

1. Introduction

1.1. Australian Englishes

In the 21st century Australian English is typically characterised as encompassing Mainstream Australian English (MAE), a group of varieties spoken by Indigenous Australians known as Australian Aboriginal Englishes (AAE), and a range of ethnocultural varieties [1]. The present paper focusses on MAE and AAE.

MAE is the most widely spoken of these varieties, and described by Cox and Fletcher [1: 11] as the "standard" and codified form used in Australia. While "standard" in this sense is not linked to correctness or prestige, attitude studies in Australia indicate that "standard" features of MAE are rated especially positively by listeners [2]. MAE is characterised as distinctive from other world Englishes by its particular combination of vowel and consonant realisations, connected speech processes, and prosodic features.

AAE varieties are spoken by First Nations people, whose ancestors spoke a wide range of Indigenous Australian languages prior to colonisation from 1788 onwards and subsequently experienced extensive language loss [e.g. 3]. AAE is an umbrella term for a range of varieties across different regions and L1 and L2 backgrounds. AAE varieties differ from MAE in various ways to varying extents at all levels of linguistic structure – in phonetics and/or phonology, grammar, semantics, pragmatics and lexicon [3-5].

1.2. Australian Englishes and the courtroom context

Originating in the 1980s [e.g. 6, 7], a growing body of work has highlighted aspects of intercultural communication (or *mis*communication) involving AAE speakers in legal processes which have critical implications for justice outcomes. Eades has written extensively on this topic and has documented a number of legal cases in which small differences in language use by AAE speakers have crucially affected the reception of evidence [e.g. 4 and papers therein, 8, 9]. Eades [4: 113] explains that while the interview format is a standard speech event in western societies, this has not been the case for Indigenous Australian societies until relatively recent times. Western-based legal processes rely heavily on interview formats: police interviews of suspects, lawyer-client interviews, and courtroom examination. She highlights a number of differences in intercultural communicative practices that can lead to misunderstandings when AAE speakers are involved in legal contexts. Among these is the observation that AAE speakers often employ and view silences in conversation in rather different ways from MAE speakers [4, see also 10], as is explored in the present study.

1.3. Australian Englishes and silences in interaction

Whereas western societies frequently see silence in conversation as negative and problematic, in Australian Indigenous communities silence can be valued positively, for example indicating that a conversation participant is thinking, or that members of the group are "enjoy[ing] the presence of others in a non-verbal way" [4: 114]. Such experiences of silence do not transfer well to contexts such as legal interviews where silence is can be interpreted as "evasion, ignorance, confusion or even guilt" [4: 114]. Eades argues that while silence cannot be taken as an admission of guilt in legal contexts, police officers, legal professionals and jurors are likely to find it difficult to ignore their cultural understanding of the meaning of silence, particularly when they are not mindful of the differing approaches to using and perceiving silences in conversation in AAE.

Several ethnographic studies comparing Australian Indigenous speakers and white Australians note differences in approaches to turn-taking and tolerance for silence [e.g. 10-14], yet there is very little empirical work examining these claims. An exception is a study by Mushin and Gardner [15] which investigates the conversation of Garrwa-speaking women, aged over 60 years, in two remote Aboriginal communities in Northern Australia. Five conversations moving between Garrwa, English and Kriol are analysed. The authors compare the length and positioning of silences in their data with previous findings on Anglo-Australian and American English using a Conversation Analysis approach. They found that silences greater than one second were more frequent than shorter silences in the Garrwa speakers' conversations, whereas the American English speakers produced short silences of 0.9-1.2s most often and far fewer longer silences. Using a range of insights from Conversation Analysis, Mushin and Gardner argue that talk among Indigenous Australian people may be qualitatively different in the way it is constructed and, that these speakers are comfortable with longer silences in their talk.

1.4. Group variation in pausing behaviour

Patterns of fluency behaviour, including pausing, can vary from one language variety to another, although few studies are available. [16] found filled pauses occurred more in Standard Southern British English than York English, while silent pauses occurred at similar rates in both varieties. [17] showed that longer silent pauses occur more frequently in York than Manchester Englishes, but that usage of other fluency features does not differ markedly across the two varieties. Among a range of differences in use of fluency features, [18] found silent pauses occurring with increasing frequency in British, American and New Zealand Englishes, in that order.

Research examining social variation in fluency features mostly focusses on filled pauses. Women have been shown to use *um* more often than *uh* [e.g. 17, 19-21]. *Um* is also more popular than *uh* with older speakers [19-21] and for speakers of higher socioeconomic status [19].

Sociolinguistic variation in fluency phenomena in Australian Englishes has received little research attention. One exception is [22] which reports a small number of differences between AAE and MAE in acoustic properties of filled pauses for speakers in the same corpora examined in the present study. Apart from Mushin and Gardner's [15] study of Garrwa women's conversation (see 1.3), variation in silent pausing in Australian Englishes is in need of quantitative investigation.

1.5. Aim

The present study investigates whether patterns of silent pausing are different across MAE and AAE varieties spoken in a non-urban location: Warrnambool, Victoria. The prediction tested is that silent pauses are more frequent and/or longer in duration in AAE compared with MAE.

2. Method

2.1. Participants

The speech of two groups of adult L1 Australian English speakers from Warrnambool, Victoria, is analysed. Warrnambool is located on the coast of south-west Victoria, approximately 250km from Melbourne. It has a population of approximately 36,000, with 3.2% of residents identifying as Aboriginal and/or Torres Straight Islander [23]. A community-driven language revitalization [see e.g. 24] is taking place for varieties of the Dhauwurd Wurrung language group in the region. The participants all identified either as male or female, and were 10 MAE speakers (4F, 6M, aged 18-72 years, mean = 33.6 years) and 10 AAE speakers (6F, 4M, aged 18-72 years, mean = 42 years). The AAE speakers included participants who identified as Gunditjmara people (Warrnambool, Framlingham) and Gunditj Mirring people (Heywood).

2.2. Data collection

The data analysed are from sociolinguistic interviews conducted by the fifth author in 2015 and 2016 as part of a larger elicitation process which also included a word list, a questionnaire and a perception study [25-27]. MAE speakers were mostly recorded in their homes in Warrnambool. AAE speakers were recorded in public spaces in Aboriginal co-operatives in Warrnambool and Heywood and in the health centre at Framlingham. Interviews ranged in duration from 3m17s to 36m45s. In the present study, 180s of "net speech" material concatenated from each speaker's interview were analysed, as yielded by interview extracts of 163s to 486s (mean extract duration = 304.2s), except in the cases of WN09 (MAE), WN25 (AAE) and WN31 (AAE) where only 125s, 167s and 123s of net speech respectively were available.

Annotation of target speech commenced at the first utterance after 60s of recording time had elapsed, to allow the interviewee time to settle in to the conversation. Once 180s net speech was reached, annotation finished at the utterance end thereafter. In cases where there was insufficient speech material after the 60s mark, the analyst returned to the first 60s of the recording to supplement the net speech collected.

2.3. Analysis of fluency using TOFFA

The present study analyses the pausing behaviour of these groups of speakers as part of a larger study looking at quantifying the broader profiles of fluency features present in the two varieties. The approach used draws on the 'Taxonomy of Fluency Features for Forensic Analysis' (TOFFA) framework, which was developed by the first and third authors to quantify fluency behaviour in naturally occurring speech for forensic phonetic purposes [28-30]. The framework categorises fluency features into the following top-level categories: Filled Pauses, Silent Pauses, Repetitions, Prolongations, Self-Interruptions. Each of these contain a number of subcategories for which full definitions are given in [28].

The present report focusses specifically on Silent Pauses; full TOFFA profiling of the speakers analysed will be reported in future work. Unlike the original version of TOFFA [28], in the present study silent pauses are not subdivided into the two subcategories of those at a grammatical boundary and those in other locations in the talk. Instead, silent pauses are subcategorised into two main types, *response latency* (RL), i.e. a silent pause immediately following a speaking turn of the interviewer (typically a question, but sometimes another type of conversational contribution), and *silent pause* (SP), which covers silent pauses occurring at any other point during the interviewee's speaking turn. The present data set includes 292 RL and 1055 SP tokens (means per speaker: 14.6 RL, 52.8 SP).

In Carroll's recent TOFFA-based study of individuals' fluency behaviour across speaking styles [17], silent pauses are subdivided via duration, as this approach proved useful in capturing individual differences between speakers. Therefore in the present study silent pauses are also subdivided according to duration, to explore whether differences between varieties may also be captured in this way. The following thresholds were used:

RL:	[rl1] ≥ 220ms	SP:	[sp1] ≥ 220ms
	[rl2] ≥ 500ms		[sp2] ≥ 500ms
	[r13] ≥ 800ms		[sp3] ≥ 800ms

Using *Praat* textgrids [31], the speech of each target speaker was transcribed orthographically. Boundaries were placed on the transcription tier at the beginning and end of each speaking turn of the target speaker to enable the calculation of the total net speech produced by the speaker for the section of the interview analysed. These net speech intervals included the duration of the RL preceding a given turn in cases where a RL was present. The full set of TOFFA fluency features were annotated on a separate tier. For the features of interest in the present study, interval boundaries were placed at the start-point and end-point of each feature. A *Praat* script was devised to extract the occurrences of the features of interest and their durations. A second *Praat* script segmented each interviewee's net speech stream into 20s time-stretches. *RStudio* was used to calculate the rate of occurrence of each fluency feature within each 20s time-stretch. Note that rates of fluency feature per unit time are used, as in more recent TOFFA work [30], rather than the syllable-base of the original paper [28].

2.4. Response latency opportunities

While capturing the rates of occurrence according to the timing thresholds of the three RL subcategories gives some insight into the differing usage of response latencies across the two varieties, it does not show the full picture. This is because the RL metrics only capture the delay between the interviewer's turn and the interviewee starting talk where there is a silence of greater than 220ms. However, participants can also respond immediately (or with less than 220ms silence), or indeed they can commence speaking before the interviewee has finished and thus create an overlap. Hence in order to gain a more nuanced view of the use of RLs, the data were revisited and an additional 'RL opportunity' tier added in the *Praat* Textgrid on which all opportunities for an RL to occur were marked, i.e. the ends of the interviewer's turns (usually a question, but sometimes a comment prompting a response). Each opportunity was classed as 'D' for delayed response, 'I' for immediate response to the interviewer's prompt (0-220ms, the threshold defining an RL) or 'O' for overlap. The total number of RL opportunities was counted and proportions of D, I and O responses of the total number of RL opportunities then calculated per speaker.

2.5. Statistical analysis

Four linear mixed effects models were calculated, two each for (a) the rate of occurrence of RLs and SPs, and (b) the duration of RLs and SPs, in *RStudio* [32] using the *lme4* package [33]. In the first set of models (1a and 1b), rate and duration were treated as continuous dependent variables, and variety (MAE and AAE) and fluency category (RL and SP) were included as the main categorical predictor variables, as well as an interaction between variety and fluency feature category. In the second set of models (2a and 2b), the model formula was the same except that fluency category was substituted for fluency subcategory (rl1, rl2, rl3; sp1, sp2, sp3) in model 2a. Comparing the durations of tokens in [rl1], [rl2], [sp1] and [sp2] subcategories across varieties would not be helpful since their membership is determined by duration. However, it is worth comparing the durations of tokens of [rl3] and [sp3] across varieties, given that there is no upper limit on the duration of tokens in these categories (≥ 800ms in both cases). Hence in model 2b, fluency category was substituted by subcategories [rl3] and [sp3]. In all models, speaker number and 20s-time-stretch were included as random intercepts. *P*-values were estimated via *t*-values using *lmerTest* [34], which approximated these values using the Satterthwaite method. A model of the same structure was calculated for each fluency category and fluency subcategory after reordering the levels for response type to allow for comparison between all variables.

For the analysis of D, O and I as a proportion of total RL opportunities, *RStudio* was used to fit beta regression models to the data using the package *betareg* [34]. The choice of model was motivated by the non-normal and non-binary distribution of the outcome variable [36]. Three models were constructed for each respective RL category (D, O and I), with variety as the categorical predictor variable. A Bonferroni correction was used due to the multiple comparisons, resulting in an adjusted significance threshold of α= 0.0163. Note that one AAE speaker (WN12) was omitted from this analysis due to essentially producing a long monologue in response to a single question for the entire section of speech analysed.

2.6. Inter-analyst consistency

For the first eight speakers, the second and third authors completed parallel but separate TOFFA analysis of the interviews. This exercise was used to calibrate the two analysts to the same standards, as well as to highlight any methodological issues and revisions to the TOFFA framework as necessary. After calibration was complete, the remaining twelve speakers were evenly assigned between the two analysts such that both completed the same number of MAE and AAE speakers.

3. Results

3.1. Rates of occurrence

Boxplots showing the distribution of speakers' rates of occurrence of RLs and SPs (with subcategories combined) for each variety are given in Figure 1.

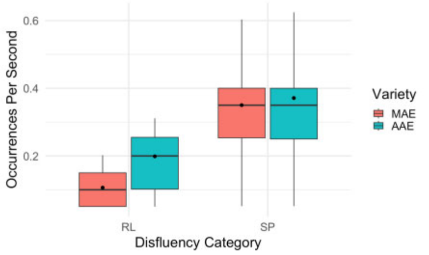


Figure 1: *Rates of occurrence of response latencies and silent pauses in MAE and AAE. Black dots (here and in subsequent figures) indicate the mean for each distribution.*

AAE speakers have a higher rate of RL (mean = 0.199 occurrences/s) than MAE speakers (mean = 0.106 occurrences/s) as is confirmed statistically (β = 0.09, *p* = 0.006). For SPs there is no significant difference between the two varieties (MAE mean = 0.350 occurrences/s, AAE mean = 0.371 occurrences/s) (β = 0.03, *p* = 0.3). In other words AAE speakers produce a long pause after a question/prompt more frequently than MAE speakers, but the two varieties display similar frequencies of occurrence of turn-internal pauses.

Boxplots showing the distribution of speakers' rates of occurrence for the three subcategories in each of RL and SP for each variety are given in Figure 2. Although [rl1] has a higher mean for AAE (0.194 occurrences/s) than MAE (0.109 occurrences/s), this is not a significant difference (β = 0.09, *p* = 0.057). However, both [rl2] and [rl3] are significantly more frequent in AAE than MAE ([rl2]: AAE

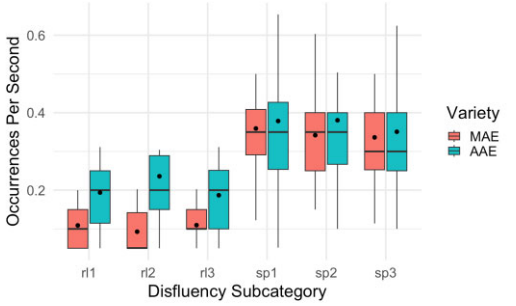


Figure 2: *Rates of occurrence of response latency subcategories* [rl1], [rl2], [rl3] *and silent pause subcategories* [sp1], [sp2], [sp3] *in MAE and AAE.*

mean = 0.236 occurrences/s, MAE mean = 0.093 occurrences/s; β = 0.12, p = 0.023; [rl3]: AAE mean = 0.187 occurrence/s, MAE mean = 0.110 occurrence/s; β = 0.09, p = 0.019). No significant differences between varieties were present for the rates of occurrence of [sp1], [sp2] and [sp3] (all $p > 0.05$).

3.2. RL opportunities

As explained in Section 2.4, in order to develop a detailed picture of the speakers' use of RLs, one must also consider how frequently RLs occur relative to the number of opportunities for an RL in a given interaction, as well as whether such opportunities were responded to with an immediate response (I), an overlap (O) or a delayed response (D). Figure 3 gives boxplots showing the distribution of speakers' proportions of D, I and O responses to RL opportunities for each variety. D responses were produced more frequently by AAE speakers (mean proportion = 0.68) than MAE speakers (mean proportion = 0.52) and this was a significant difference (z = -2.454, p = 0.0141). I and O responses were produced less frequently by AAE speakers (I mean = 0.15, O mean = 0.18) than by MAE speakers (I mean = 0.22, O mean = 0.26). However, neither of these differences were significant (for I, z = -1.478, p = 0.139; for O, z = -1.551, p = 0.121). In other words, delayed responses were used more frequently by AAE speakers than MAE speakers, while MAE speakers showed a descriptive but not statistical tendency to speak immediately or even overlap with the interlocutor more frequently than AAE speakers.

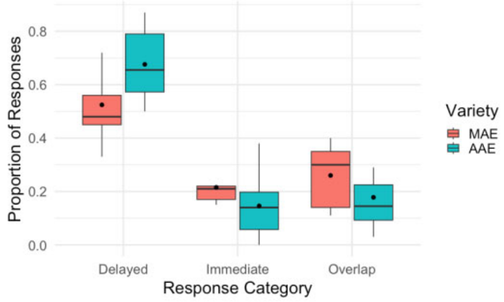


Figure 3: *Boxplots of proportions per speaker of response latency opportunities responded to with a delayed response (D), an immediate response (I) or an overlap (O).*

3.3. Duration

For each variety, boxplots showing the distribution of the duration of RLs and SPs (with durational subcategories combined) are in Figure 4. While RL duration is descriptively slightly higher in AAE (mean = 1.154s) than MAE (mean = 1.003s), there is no statistically significant difference (p = 0.08). For SP durations, there is no significant difference between varieties (MAE mean = 0.610s, AAE mean = 0.704) (p = 0.11). For [rl3], AAE speakers (mean = 1.701s) have descriptively longer RLs than MAE speakers (mean = 1.432s), but this difference is not statistically significant. Similarly, [sp3] has descriptively longer durations for AAE speakers (mean = 1.308s) than for MAE speakers (mean = 1.247s), yet this difference also does not reach statistical significance.

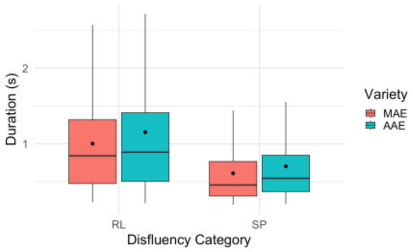


Figure 4: *Boxplots of durations of response latencies and silent pauses in MAE and AAE.*

4. Discussion and Conclusion

The study's prediction that silences would be used more frequently in AAE than MAE was confirmed for RLs, but not turn-internal SPs where no differences between varieties were observed. These RLs were descriptively, but not statistically longer for AAE speakers. Analysing RL opportunities showed that AAE speakers responded statistically more frequently with a delayed response than MAE speakers. MAE speakers produced immediate responses and overlap responses descriptively more often than AAE speakers. These findings are consistent with the qualitative observations of Eades and other researchers that AAE speakers use silent pausing in different ways in the structure of conversation. Such differences could have crucial implications in legal interview contexts [e.g. 4].

In Australia, the majority of professionals conducting legal interviews and court proceedings such as police officers, judges and lawyers, are likely to speak MAE. Indeed, the majority of people witnessing such speech events, including jury members, are likely to be MAE speakers. An AAE speaker's tendency not to respond promptly to a question in these contexts has the potential to be misinterpreted, for example as hedging or evasion. Further, as pointed out by Eades [4], an AAE speaker's answer to a question may actually be interrupted due to an interrogator's misunderstanding of ways of using silence in these varieties: "if we accept that the first part of an Aboriginal answer often starts with silence, then to start the next question before the Aboriginal interviewee has had the time to speak is in effect to interrupt the first part of the answer" [4: 178]. The present study's results provide quantitative evidence justifying the concerns of Eades and other researchers that cultural differences in interpreting silences could have a serious impact on how AAE-speaking witness's stories are told and received in legal settings. Future research should explore this further by extending the quantitative approach demonstrated in the present work to speech data from genuine police and courtroom contexts and to other sociolinguistically relevant contexts such as dialogue between two AAE speakers.

The findings also offer further confirmation that speakers of different language varieties use fluency features in different ways [cf. 16-18]. The present study focused on RLs and SPs, but it will be important to consider in future work how other fluency features behave alongside silences, especially filled pauses and prolongations of speech sounds. All of these fluency features have a range of purposes and contribute to conversational management and speech planning work in differing ways [e.g. 37-39]. Work such as [16, 28, 29] shows that an increase in use of one fluency feature can be accompanied by a decrease in the use of another both at an individual and group level. Further work underway will develop full fluency profiles for these AAE and MAE speakers to determine the extent of individual and group variation present among the different fluency features and their interrelationships. The present dataset was too small to enable analysis of age and gender patterns; these should be examined with a larger dataset in future work.

5. Abbreviations

AAE	Australian Aboriginal Englishes
MAE	Mainstream Australian English
RL	Response latency
[rl1], [rl2], [rl3]	RL subcategories – see 2.3
SP	Silent pause
[sp1], [sp2], [sp3]	SP subcategories – see 2.3
D	RL opportunity responded to with a delayed response
I	RL opportunity responded to with an immediate response
O	RL opportunity responded to with an overlap
TOFFA	Taxonomy of Fluency Features for Forensic Analysis

6. Acknowledgements

This research is supported by a Cambridge Humanities Research Grant 'Sociophonetic Variation in Aboriginal and Mainstream Australian Englishes: A Study of Disfluency Features.

7. References

[1] Cox, F. and Fletcher, J., Australian English: Pronunciation and Transcription. Cambridge University Press, 2017.

[2] Willoughby, L., "Attitudes to Australian English", in L. Willoughby and H. Manns [Eds], Australian English Reimagined: Structure, Features and Developments, 224-237, Routledge, 2020.

[3] Dickson, G., "Aboriginal English(es)", in L. Willoughby and H. Manns [Eds], Australian English Reimagined: Structure, Features and Developments, 134-154, Routledge, 2020.

[4] Eades, D., Aboriginal Ways of Using English. Aboriginal Studies Press, 2013.

[5] Malcolm, I. G., Australian Aboriginal English: Change and Continuity in an Adopted Language. De Gruyter Mouton, 2018.

[6] Koch, H., "Nonstandard English in an Aboriginal land claim", in J. Pride [Ed], Crosscultural Encounters: Communication and Miscommunication, 176-195, River Seine Publications, 1985.

[7] Eades, D., "Sociolinguistic evidence in court", Australian Journal of Communication, 14:22-33, 1988.

[8] Eades, D., "The social consequences of language ideologies in courtroom cross-examination", Language in Society, 419(4):471-497, 2012.

[9] Eades, D., "Communication with Aboriginal Speakers of English in the legal process", Australian Journal of Linguistics, 32(4):473-489, 2012.

[10] Eades, D., "Understanding Aboriginal silence in legal contexts", in H. Kotthoff and H. Spencer-Oatey [Eds], Handbook of Intercultural Communication, 285-301, Mouton de Gruyter, 2007.

[11] Eades, D., "Communicative strategies in Aboriginal English", in S. Romaine [Ed], Language in Australia, 84-93, Cambridge University Press, 1991.

[12] Eades, D., "I don't think it's an answer to the question: Silencing Aboriginal witnesses in court", Language in Society, 29(2):161-195, 2000.

[13] Liberman, K., Understanding Interaction in Central Australia: An Ethnomethodological Study of Australian Aboriginal People. Routledge and Kegan Paul, 1985.

[14] Walsh, M., "Interactional styles in the courtroom: an example from northern Australia", in J. Gibbons [Ed], Language and the Law, 217-233, Longman, 1994.

[15] Mushin, I. and Gardner, R., "Silence is talk: Conversational silence in Australian Aboriginal talk-in-interaction", Journal of Pragmatics, 419(10):2033-2052, 2009.

[16] McDougall, K., Duckworth, M., and Hudson, T., "Individual and group variation in disfluency features: a cross-accent investigation", Proc. 18th ICPhS, Paper number 0308, 1-5.

[17] Carroll, L., A Forensic Phonetic Investigation of Disfluency Behaviour across Two Interactional Styles: Applying a Modified Analytical Framework TOFFAMo, M.A. Dissertation, University of Lancaster, 2019.

[18] Shanmukha, N., A Corpus-Based Study of Speech Fluency across English Dialects, M.Sc. Dissertation, University of Canterbury, 2017.

[19] Tottie, G., "*Uh* and *Um* as sociolinguistic markers in British English", International Journal of Corpus Linguistics, 16(2):173-197, 2011.

[20] Acton, E., "On gender differences in the distribution of *um* and *uh*", University of Pennsylvania Working Papers in Linguistics, 17(2), Article 2, 2011.

[21] Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J. and Liberman, M., "Variation and change in the use of hesitation markers in Germanic languages", Language Dynamics and Change, 6(2):199-234, 2016.

[22] Blackwell, L. and McDougall, K. "Sociophonetic variation of filled pauses in Victoria, Australia", Proc. 19th SST, Melbourne, this volume, 2024.

[23] Australian Bureau of Statistics, "Census Quick Stats", 2021.

[24] Eira, C., "Addressing the Ground of Language Endangerment", Working together for Endangered Languages: Research Challenges and Social Impacts – Proceedings of Foundation for Endangered Languages Conference XI, 82-90, 2007. https://vacl.org.au/wp-content/uploads/2021/12/addressing-the-ground-of-language-endang.pdf

[25] Loakes, D., McDougall, K., Clothier, J., Hajek, J. and Fletcher, J., "Sociophonetic variability of post-vocalic /t/ in Aboriginal and mainstream Australian English", Proc. 17th SST, 5-8, 2018.

[26] Loakes, D., McDougall, K., and Gregory, A., "Variation in /t/ in Aboriginal and Mainstream Australian Englishes", Proc. 18th SST, Canberra, 2022, 61-65.

[27] Loakes, D. and Gregory, A., "Voice quality in Australian English", JASA Express Letters, 2(08/01):085201, 2022.

[28] McDougall, K. and Duckworth, M., "Profiling fluency: an analysis of individual variation in disfluencies in adult males", Speech Communication, 95:16-27, 2017.

[29] McDougall, K. and Duckworth, M., "Individual patterns of disfluency across speaking styles: a forensic phonetic investigation of Standard Southern British English", Int. J. Speech Lang. Law, 25(2):205-230, 2018.

[30] McDougall, K., Rhodes, R., Duckworth, M., French, J. P. and Kirchhübel, C., "Application of the 'TOFFA' framework to the analysis of disfluencies in forensic phonetic casework", in Proc. 18th ICPhS, 731-735.

[31] *Praat*: Doing phonetics by computer. [Computer program]. (1992-2024). Available: http://www.praat.org/.

[32] R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, 2024. Available: https://www.R-project.org/

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

[34] Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B., "*lmerTest* package: tests in linear mixed effects models", Journal of Statistical Software, 82(13):1-26, 2017.

[35] Cribari-Neto, F. and Zeileis, A., "Beta regression in R", Journal of Statistical Software, 34(2):1-24, 2010.

[36] Ferrari, S. L. P. and Cribari-Neto, F., "Beta regression for modeling rates and proportions", Journal of Applied Statistics, 31(7):799-815, 2004.

[37] Shriberg, E., "To 'errrr' is human: ecology and acoustics of speech disfluencies", Journal of the International Phonetic Association, 31(1):153-169, 2001.

[38] Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E., "Disfluency rates in conversation: effects of age, relationship, topic, role and gender", Language and Speech 44(2):123-147, 2001.

[39] MacGregor, L. J., Corley, M. and Donaldson, D. I., "Listening to the sound of silence: disfluent silent pauses in speech have consequences for listeners", Neuropsychologia, 48(14):3982-3992, 2010.

Sociophonetic Variation of Filled Pauses in Victoria, Australia

Liz Blackwell, Kirsty McDougall

University of Cambridge

ecb76@cam.ac.uk; kem37@cam.ac.uk

Abstract

An acoustic study of sociophonetic variation in the fillers ‘um’ and ‘uh’ in Australian Englishes (Aboriginal, AAE; Mainstream, MAE) spoken in Warrnambool, Victoria, is presented. Duration, f0 and formant frequency measurements of fillers produced by 14 AAE and 14 MAE speakers were analysed. ‘uh’ and ‘um’ exhibited different vowel qualities which also varied with gender, variety and age. ‘um’ demonstrated a longer vowel duration than ‘uh’ for male MAE speakers only. Female MAE speakers produced ‘um’ with a higher f0 than female AAE speakers; no such pattern was present for male speakers, nor for ‘uh’ produced by any speaker group.

Index Terms: fillers, filled pauses, fluency features, Australian Englishes, Australian Aboriginal Englishes

1. Introduction

1.1. Australian Englishes

At least 200 languages were spoken in Australia before European colonisation [1, 2], but following colonial suppression of First Nations culture and language and the ensuing dominance of English, there now remain an estimated 20 Indigenous languages, half of which are spoken by fewer than 1000 speakers [3]. A typology of contemporary Australian Englishes [4] suggests a broad tripartite division into the supra-regional standard (‘Mainstream’ Australian English, MAE), Aboriginal Australian Englishes (AAEs) and, outside the scope of the present paper, Ethnocultural Australian Englishes such as Greek or Yiddish Australian English [5].

Within MAE and AAE, there is a significant amount of social and geographical variation. MAE has historically been described as encompassing variation between ‘broad’ and ‘cultivated’ sociolects, but this is now considered outdated, and suggested to have been replaced by a new “general” prestige variety, with social and regional variation undergoing reorganisation in accordance with Australia’s developing post-colonial identity [6]. Regional variation is subtle, but increasingly noted in phonetic studies [7, 8]

AAEs are a distinct group of Englishes that exhibit variation from MAE at many levels of linguistic structure, including phonology and lexicon [3] as well as notable differences in pragmatic and discoursal features [3, 9, 10]. It has been noted that AAEs may differ from MAE with respect to filler words and phrases such as “might be” or “like” [3], but ‘um’ and ‘uh’ are not mentioned in these observations. Additionally, crucial differences in the use of silence and pausing behaviour in AAEs, namely a positive view of silence and pausing in conversation by Indigenous speakers, have been documented in courtroom contexts [9]. Here, “silence fillers” such as ‘um’ and ‘uh’ are suggested to be a strategy employed more in western cultures, though a lower rate of filler use by Indigenous speakers is not directly demonstrated.

1.2. Social variation in ‘um’ and ‘uh’ distribution

Existing studies of social variation in filler use have focused on non-acoustic variables such as relative proportion and frequency. Frequency has been shown to correlate with both gender and age: in American English, men have repeatedly demonstrated a higher frequency of filler usage [11, 12], and increasing age correlates with more frequent filler use [12]. The relative proportion of ‘um’ and ‘uh’ has been shown to correlate with age and gender in American English [13], with male speakers and older speakers having a higher relative proportion of ‘uh’ than their female or younger counterparts. The same has been shown to be true in British English [14], together with a higher overall filler frequency in male speech, and greater relative use of ‘uh’ by both younger male speakers and older speakers of any gender.

Clearly, fillers may vary along social lines. However, from a forensic perspective, filler frequency has also been shown to be highly speaker-specific within a socially homogeneous group [15], with rate of ‘um’ production allowing for discrimination between male Southern Standard British English (SSBE) speakers at a level three times greater than chance.

1.3. Acoustic parameters of ‘um’ and ‘uh’

The vowel quality of fillers differs across languages [16, 17], even within bilingual speakers when switching language [16]. Filler duration has also been shown to be cross-linguistically variable, with French fillers being demonstrably longer than their German equivalents [16]. However, there is little research available on these factors at the sociolinguistic level. In forensic research, acoustic properties of ‘um’ and ‘uh’ have been shown to be comparable to lexical vowels in their capacity for speaker differentiation [18], using F1-F3 formant frequencies, vowel duration, and nasal duration in ‘um’. Whilst both ‘um’ and ‘uh’ were investigated, ‘um’ was found to be a greater locus of individual variation, whereas ‘uh’ was more consistent across speakers, all of whom were young, male SSBE speakers [19]. Differences in acoustic features of ‘um’ and ‘uh’ have also been reported. In French and German [16], and again in SSBE [18], the vocalic portion of ‘um’ is consistently shorter than the duration of ‘uh’, though ‘um’ is the longer of the two when the nasal portion is included. There have been varying conclusions on whether ‘uh’ and ‘um’ have different vowel qualities. F2 frequency has been reported significantly higher for ‘uh’ in German, and F1 frequency significantly higher for ‘um’ in both French and German [16].

However in British English, static formant frequency measurements of ‘um’ and ‘uh’ have been suggested to differ very little [18].

1.4. Aims

Social variation has been demonstrated in filler choice and their frequency of occurrence, and trends appear consistent across British and American English. However to date, acoustic variation in filler realisation has only been shown across languages, or between individuals, rather than socially delimited. The present study therefore examines a group of speakers from Warrnambool, Victoria, investigating whether the factors of gender, age, or variety of Australian English (MAE or AAE), correlate with the variables of duration, f0, and F1-F2 formant frequencies.

2. Method

2.1. Speakers

The speech analysed in the present study forms part of a set of sociolinguistic interviews conducted in 2015 and 2016 by Debbie Loakes at the University of Melbourne. The subset of data used in this study represents informal, spontaneous speech from 29 residents of Warrnambool, a regional centre in south-western Victoria. Warrnambool, around 3 hours’ travel from Melbourne, is a small city of 35,000 residents, of which 2% - double the state average - identify as Indigenous [20]. One male MAE speaker, who produced no clear fillers in the course of his interview, was excluded from further analysis. Of the 28 remaining speakers, 14 are speakers of AAE (9 female, 5 male), and 14 speakers of MAE (7 female, 7 male). Ages range from 18 to 72 years, with a mean of 36; however, the mean age of the MAE speakers is 42, in contrast to 30 for the AAE speakers. Length of speech material from each specific participant varied greatly, ranging from 47 seconds to 30 minutes, with a median of 3 minutes and 46 seconds.

2.2. Data collection

Transcription was carried out in *Praat* [21], to tiered textgrids. The first tier was used to indicate when the target speaker began and ceased speech, and filler tokens were marked for more detailed transcription on the second tier. Following the process of [18], tokens of ‘um’ were split into vocalic and nasal portions, where visual inspection of the spectrogram mostly showed a clear nasal onset with sharp decreases in the amplitudes of formant frequencies. Where not spectrographically visible, onsets were marked using auditory analysis. After discarding any tokens in overlapping speech, 1204 filler tokens were available for analysis, with 310 tokens of ‘uh’ and 894 tokens of ‘um’.

2.4. Analysis

Duration, fundamental frequency (f0), and midpoint F1-F2 frequency values were extracted using *Praat* scripting. For ‘um’, measurements were taken at the midpoint of the vocalic portion. For the purposes of this paper, the vowel in both ‘um’ and ‘uh’ was taken to be a short vowel, /ɐ/. Whilst the presence of nasalisation in ‘um’ may also affect vowel quality, in SSBE the coarticulatory effect of /m/ in ‘um’ has been shown to only begin around 70%-90% of the way through the /ɐ/ vowel [18], and so static midpoint measurements are less likely to demonstrate any significant effect from nasal coarticulation. Tokens returning ‘undefined’ or outlier f0 values were checked and discarded if found to have non-modal voice quality, removing 146 tokens from further analysis, leaving a total sample of 1058 tokens. Over half of these discarded tokens were from a single speaker with a particularly ‘creaky’ speech style, but over 100 clear tokens from this speaker remained for analysis. Of the 1058 tokens available for further analysis, 543 were produced by female speakers (53 ‘uh’, 489 ‘um’), and 515 by male speakers (188 ‘uh’, 344 ‘um’). However, due to differences in interview lengths, there was an uneven distribution of fillers between dialect groups, with 708 tokens by MAE speakers (193 uh, 531 um) and 350 tokens by AAE speakers (48 uh, 302 um) available for analysis.

Data were fitted to a series of linear mixed-effects models in *RStudio* [22], using the *lme4* package [23]. Dialect, gender, age and filler type were each used as categorical predictors, with speakers divided into two age groups of <40 years and >40 years. For models including f0 and F1-F2 formant frequencies, analysis of male and female speakers was carried out separately. A random intercept for speaker ID was also included in each model. The *RStudio* package *gtsummary* [24] was then used to estimate *p*-values resulting from the *t*-values in the output produced by the models.

3. Results

3.1. Duration

For the group at large, the vowel of ‘um’ was shorter than that of ‘uh’ (‘um’ mean = 0.29*s*, ‘uh’ mean = 0.33*s*, $\beta =- 0.02s,\ p < 0.038$), and the total length of ‘um’ was significantly longer than that of ‘uh’ (‘um’ mean = 0.48*s*, ‘uh’ mean = 0.33*s*, $\beta = 0.19s,\ p < 0.001$). However, male MAE speakers had noticeably different results: the vowel in ‘um’ produced by these speakers was significantly *longer* than that of ‘uh’ (‘um’ mean = 0.36*s*, ‘uh’ mean = 0.30*s*, $\beta = 0.04s,\ p < 0.001$), leading to a greater durational difference between the two fillers for male MAE speakers.

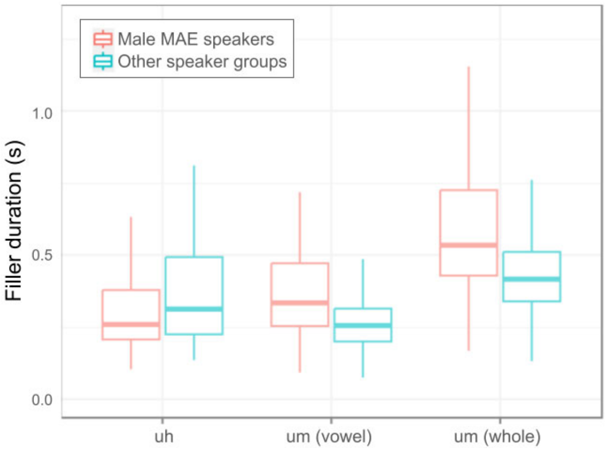


Figure 1: *Relative durations of ‘uh’ and ‘um’.*

3.2. F0

'Um' produced by female speakers showed a significant dialect effect, with female speakers of MAE producing 'um' with a higher average f0 than female speakers of AAE (MAE mean = 181*Hz*, AAE mean = 156*Hz*, $\beta = 47Hz,\ p = 0.009$).

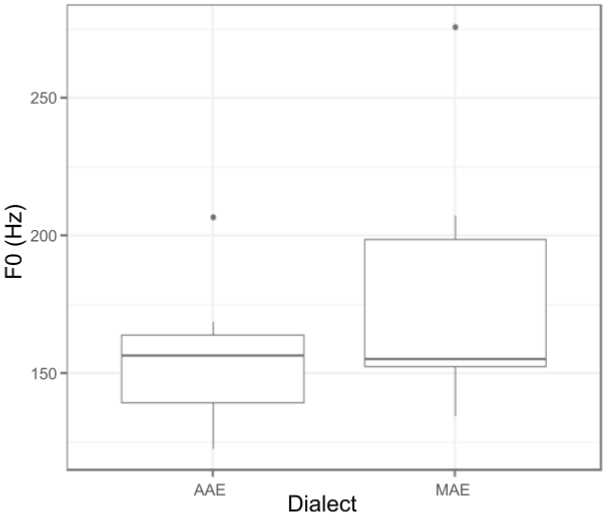


Figure 2: *F0 of 'um' by female speakers.*

There was no such interaction for 'uh' by female speakers, nor any such interaction for either filler in the male speaker group.

3.3. F1 and F2 formant frequencies

F2 of 'uh' in female MAE speakers was significantly higher in frequency than 'um', corresponding to a 'fronted' vowel quality ('um' mean = 821*Hz*, 'uh' mean = 784*Hz*, $\beta = 113Hz,\ p < 0.001$).

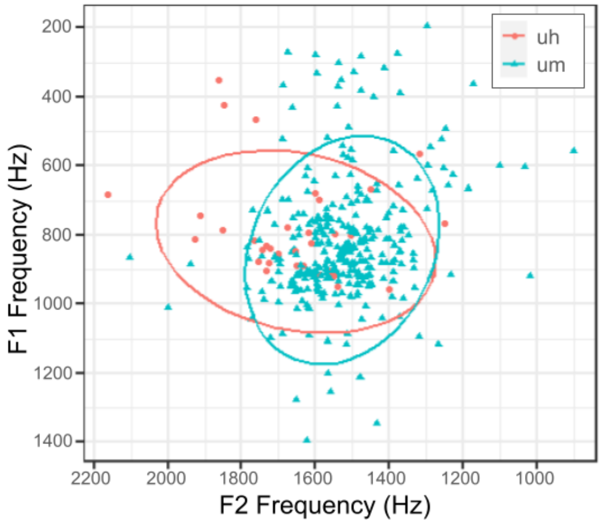


Figure 3: *95% confidence ellipses for 'uh' and 'um' produced by female MAE speakers.*

F1 of 'uh' produced by female MAE speakers under 40 was much higher than 'uh' in those over 40 (<40 mean = 851*Hz*, >40 mean = 698*Hz*, $\beta = 153Hz,\ p < 0.001$), corresponding to a lowered 'uh' vowel in younger female speakers.

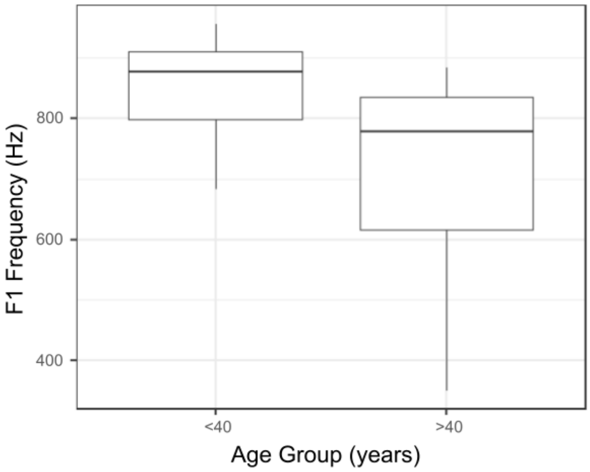


Figure 4: *F1 frequency of 'uh' by female MAE speakers by age group.*

Midpoint F1 and F2 frequencies of individual tokens of 'uh' and 'um' produced by male speakers of MAE are shown in Figure 5, with 95% confidence ellipses. Male speakers tended to have a higher F1 frequency in 'uh' than 'um', a trend reaching statistical significance for MAE speakers ('uh' mean = 598*Hz*, 'um' mean = 616*Hz*, $\beta = 27Hz,\ p = 0.005$), but not AAE speakers ($\beta = 39Hz,\ p = 0.065$).

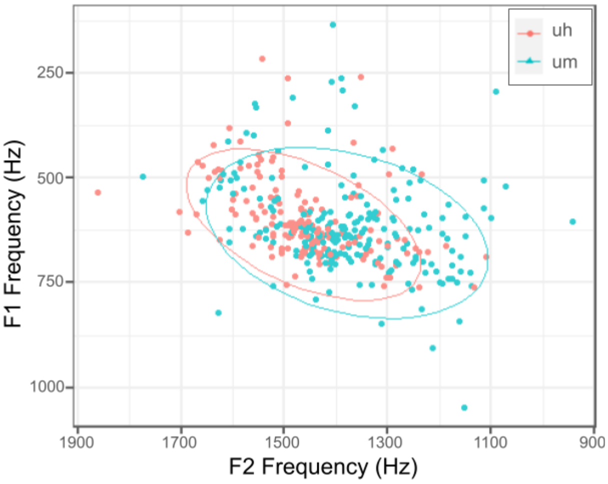


Figure 5: *95% confidence ellipses for F1 and F2 frequencies of 'uh' and 'um' produced by male speakers of MAE.*

4. Discussion

4.1. Relative duration findings

As discussed in section 1.3, previous studies of relative duration in 'um' and 'uh' [16, 18] have consistently shown the vowel in 'um' to be shorter than 'uh'. However, the findings in section 3.1 show that although AAE speakers and female MAE speakers maintain this relationship, male MAE speakers produce 'um' with a longer vowel than 'uh', presenting the possibility that relative duration may vary according to dialectal factors. Additionally, both [18] and [16] worked with single-sex samples, and so the present study demonstrates the potential importance of gender as a factor in filler duration. A major factor affecting segmental duration is placement within the intonational phrase (IP) [25], a factor not considered in the present study. Therefore, further explanation of this finding would require examination of the positions of 'uh' and 'um' within the IP. Furthermore, participants' individual speech rates were not calculated, but no significant gender, dialect or age differences in filler frequency per minute were found [26].

4.2. F0 findings

Previous research on f0 and Warrnambool dialect groups has suggested a lower average f0 for male AAE speakers [27]. However, the present study did not find any such trend for male speakers, and f0 only displayed a significant drop for 'um' in female AAE speakers. As 'uh' displayed no such trend ($p = 0.2$), differences in prosodic placement once again present a potential factor, requiring further investigation. The lack of any dialect trend in f0 concurrent with wider literature is possibly due to the restricted style and content of speech in the present study.

4.3. F1 and F2 formant frequency findings

4.3.1. Differences in placement of 'uh' and 'um'

As discussed in section 1.3, differences in vowel quality between 'um' and 'uh' have been reported in French and German [16], but were not found in SSBE [18]. However, the present study provides preliminary evidence that 'uh' and 'um' may have significant, if slight, differences in vowel placement. Based on findings for F1 (section 3.3), male MAE speakers produced 'uh' with a slightly higher vowel quality than 'um', a trend potentially present, but below the level of significance, for male AAE speakers. In contrast, findings for F2 showed that for female MAE speakers, 'uh' had a more fronted quality than 'um' (Fig. 3). Therefore, 'uh' and 'um' may differ in vowel placement in some dialects of English, and this placement may also be sensitive to social factors such as gender. However, if onset of nasalisation in AusE is comparatively earlier than that of SSBE, where /ɐ/ in 'um' typically begins to nasalise in the final third of the vowel [18] then it is also possible that nasal coarticulation had an effect on the midpoint formant frequencies.

4.3.2. Age effect and potential connection to short front vowel lowering

An age effect was also present for female MAE speakers, with younger participants producing 'uh' with much lower vowel quality (Fig. 4). No such trend was demonstrated for speakers of AAE, or for 'um'. This finding potentially relates to the ongoing change of short front vowel lowering found in Sydney speakers of MAE [28], a change that is not particularly evident in the speech of AAE speakers in Warrnambool [29]. The exclusivity of the trend to 'uh', and not 'um', may relate to the fronting of 'uh' by these speakers as discussed in section 4.3.1, or potentially may be due to the apparent resistance of nasalised vowels to the change, as previously demonstrated for nasalised [æ̃] [30]. However, the small sample size of 'uh' from female MAE speakers (n = 33 tokens) may have also been a contributing factor, so further data are necessary for any comprehensive investigation.

5. Conclusions

This paper has presented a preliminary investigation of sociophonetic variation in filled pauses produced by speakers of two varieties of Australian English, MAE and AAE. Gender, dialect and age were shown to be potential factors of variation in vowel duration, f0, and vowel placement. Findings also support the treatment of 'uh' and 'um' as separate entities when investigating vowel quality of fillers, as vowel quality differed between fillers, and relative differences in vowel quality were also shown to be socially variable. The vowel duration of 'um' was longer than 'uh' in male MAE speakers, a novel finding that contrasts with previous studies of British English, French and German. Whilst a tentative connection was made between short front vowel lowering and the lowered quality of 'uh' for some speakers, further research and a larger dataset are necessary for any strong conclusion. For any further research regarding social variation in duration and f0, investigation of the relationship between social factors and prosodic placement of fillers is also needed.

6. Acknowledgements

The authors are grateful to Debbie Loakes, The University of Melbourne, for making the data set available for this study.

7. References

[1] Malcolm, I. G., "Aboriginal English research: an overview", Asian Englishes, 3(2):9–31, 2000.

[2] Burridge, K., "History of Australian English", in L. Willoughby and H. Manns [Eds], Australian English Reimagined, 175–192, Routledge, 2019.

[3] Butcher, A., "Linguistic aspects of Australian Aboriginal English", Clinical Linguistics & Phonetics, 22(8):625–642, 2008.

[4] Cox, F. and Fletcher, J., Australian English Pronunciation and Transcription, Cambridge University Press, 2017.

[5] Clyne, M., Eisikovits, E., and Tollfree, L. "Ethnic varieties of Australian English", in D. Blair and P. Collins [Eds], English in Australia, 223–238, John Benjamins, 2001.

[6] Cox, F. and Palethorpe, S., "Standard Australian English: the sociolinguistic broadness continuum", in R. Hickey [Ed], Standards of English, 294–317, 2012.

[7] Cox, F. and Palethorpe, S., "Vowel variation across four major Australian cities", in S. Calhoun, P. Escudero, M. Tabain and P. Warren [Eds], Proc. 19th International Congress of Phonetic Sciences 2019, 577–581, ASSTA, 2019.

[8] Loakes, D., Clothier, J., Hajek, J. and Fletcher, J., "An investigation of the /el/-/æl/ merger in Australian English: a pilot study on production and perception in South-West Victoria", Australian Journal of Linguistics, 34(4):436–452, 2014.

[9] Eades, D., "Understanding Aboriginal silence in legal contexts", in H. Kotthoff and H. Spencer-Oatey [Eds], Handbook of Intercultural Communication, 285–301, Mouton de Gruyter, 2007.

[10] McDougall, K., Paver, A., Duckworth, M., Blackwell, L. and Loakes, D. "Patterns of silent pausing in Aboriginal and

Table of Contents for this manuscript

Mainstream Australian Englishes spoken in Warrnambool", Proc. 19th Australasian International Conference on Speech Science and Technology, Melbourne, this volume, 2024.

[11] Shriberg, E., "Disfluencies in Switchboard", Proc. International Conference on Spoken Language Processing, 11–14, Philadelphia, PA: IEEE, 1996.

[12] Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F. and Brennan, S. E., "Disfluency rates in conversation: effects of age, relationship, topic, role, and gender", Language and Speech, 44(2):123–147, 2001.

[13] Acton, E. K., "On gender differences in the distribution of um and uh", University of Pennsylvania Working Papers in Linguistics, 17(2):2, 2011.

[14] Tottie, G., "Uh and um as sociolinguistic markers in British English", International Journal of Corpus Linguistics, 16(2):173–197, 2011.

[15] McDougall, K. and Duckworth, M. "Profiling fluency: an analysis of individual variation in disfluencies in adult males", Speech Communication, 95:16–27, 2017.

[16] Lo, J. J. H., "Between äh(m) and euh(m): the distribution and realization of filled pauses in the speech of German-French simultaneous bilinguals", Language and Speech, 63(4):746–768, 2020.

[17] Candea, M., Vasilescu, I. and Adda-Decker, M., "Inter- and intra-language acoustic analysis of autonomous fillers", Proc. Disfluency in Spontaneous Speech Workshop, 47–51, Aix-en-Provence, 2005.

[18] Hughes, V., Wood, S. and Foulkes, P., "Strength of forensic voice comparison evidence from the acoustics of filled pauses", International Journal of Speech, Language and the Law, 23(1):99–132, 2016.

[19] Nolan, F., McDougall, K., de Long, G. and Hudson, T., "The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research", International Journal of Speech, Language and the Law, 16(1):31–57, 2009.

[20] Australian Bureau of Statistics, "2021 Census All persons QuickStats", 2021. Online: https://www.abs.gov.au/census/find-census-data/quickstats/2021/2021, accessed on 07 Jul 2024.

[21] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", 2024. Online: http://www.praat.org/.

[22] Posit Team, "RStudio: integrated development environment for R." Posit Software, PBC, Boston, MA, 2023. Online: http://www.posit.co/.

[23] Bates, D., Mächler, M., Bolker, B. and Walker, S., "Fitting linear mixed-effects models using lme4.", Journal of Statistical Software, 67(1):1–48, 2015.

[24] Sjoberg, D. D., Whiting, K., Curry, M., Lavery, J. A. and Larmarange, J., "Reproducible summary tables with the gtsummary package", The R Journal, 13(1):570–580, 2021.

[25] Nolan, F., "Intonation", in B. Aarts and A. McMahon [Eds], The Handbook of English Linguistics, 433–457, Blackwell, 2006.

[26] Blackwell, L., "A study of sociophonetic variability in filled pauses in Southeastern Australia", Undergraduate dissertation, University of Cambridge, 2024.

[27] Loakes, D. and Gregory, A., "Voice quality in Australian English", Journal of the Acoustical Society of America Express Letters, 2(8):085201, 2022.

[28] Cox, F. and Palethorpe, S., "Reversal of Short Front Vowel Raising in Australian English", Proc. Annual Conference of the International Speech Communication Association, 342–345, Interspeech, 2008.

[29] Loakes, D., Fletcher, J., Hajek, J., Clothier, J. and Volchok, J. "Short vowels in L1 Aboriginal English spoken in Western Victoria.", in C. Carignan and M. Tyler [Eds], Proc. 16th Australasian International Conference on Speech Science and Technology, 33–36, ASSTA, 2016.

[30] Cox F. and Palethorpe, S., "Phonologisation of vowel duration and nasalised /æ/ in Australian English", in J. Hay and E. Parnell [Eds], Proc. 16th Australasian International Conference on Speech Science and Technology, 33–36, ASSTA, 2014.

Masculinity and Sexual Orientation as Predictors of Creaky Voice in Australian English Speaking Men

Timothy Shea, Hannah White, Joshua Penney, Felicity Cox

Department of Linguistics, Macquarie University
timothy.shea@hdr.mq.edu.au; hannah.white, joshua.penney, felicity.cox@mq.edu.au

Abstract

Pop culture strongly associates creaky voice with young women and, perhaps less strongly, with gay men. However, empirical investigations into variation in creak by gender have produced equivocal findings. This study sought to examine the prevalence of creak in male speech and contribute to the nascent literature covering Australian English (AusE) variation by sexual orientation (SO). Intervals of creak in speech data from male AusE speakers were considered with regard to speaker age, SO, orientation toward traditional male gender roles, and speech topic. Results revealed a complex interaction effect between the first three of these four variables on creak prevalence.

Index Terms: sociophonetics, sexual orientation, masculinity, creaky voice, Australian English

1. Introduction

1.1. Gay Men's Speech

Over the past three decades, much sociophonetic research has investigated the speech of gay men, with early studies seeking to empirically identify the phonetic variants that might index male SO. Fundamental frequency (f0) related metrics have often featured in these studies, but research has failed to produce a consensus as to whether and how f0 may systematically vary according to speaker SO. Further, the way in which gay male speech has been conceptualised has changed throughout that period. Early studies treated gay male speech as a monolithic category and sought to directly link it to specific phonetic variants ([1, 2]). Subsequent studies have taken a more nuanced approach, examining variation within and between speech communities of gay people in specific geographic locations [3]. Attention has also turned to the ways in which gay and other sexual minority [4] (i.e., not straight) speakers might actively employ phonetic variants to specifically access the social meanings that they index, and in doing so construe and communicate their identities [5].

The vast majority of this work has focused on Northern Hemisphere varieties of English ([1, 2]) and other, mostly European, languages ([6, 7]), although some has examined variation in New Zealand English [8]. However, there remains a gap in scholarly understanding of the phonetic features of gay men's speech in the AusE context (but see [9] regarding perceived-as-gay speech and [10] regarding the acoustics of /s/).

1.2. Gender and Masculinity

Hegemonic masculinity describes a combination of social influences and actions which serve both to dominate women and to subordinate other alternative masculinities [11]. In doing so, it privileges the minority of men who embody its ideals, which include toughness, stoicism, and heterosexuality. Being a hegemony, it polices the ways men are expected to behave, and traits they are expected to adopt, vis-à-vis the performance of gender [12]. The list of these policed attributes is extensive. For example, in addition to avoiding styles of dress and grooming associated with women, in relationships men are expected to adhere to defined gender roles which project dominance and hide vulnerability. In addition, the pressures of hegemonic masculinity dictate that men not deviate from a prescribed range of masculine speech styles [13] under penalty of being labelled unmanly or womanly.

Among the phonetic variables which might situate a speaker inside (or outside) of these licensed speech styles is variation related to f0. For example, in American English, falling prosodic contours and lower pitch ([14, 15]) have been linked to sounding masculine. Additionally, male voices and phonetic variants that are perceived as gay are regularly also interpreted as less masculine ([14, 15, 16]).

Openly gay men, as well as other men who openly belong to the sexual minority, necessarily deviate from hegemonic masculinity. It might therefore be the case that they are less likely to be influenced by the pressure to restrict their speech styles than straight men are. This cannot be taken for granted, however, particularly in light of the aforementioned lack of consensus regarding f0 measurements in gay male speech [1]. We ought to consider the possibility that the effect of differing speaker orientations to masculinity contributed to these studies' mixture of findings. Furthermore, given that gay men are already situated outside, or on the borders of, the masculine hegemony, if masculinity does impact the ways they use f0, it may do so in ways that differ from their straight counterparts.

While gender associations for a variable such as pitch are intuitive because the typically larger and wider vocal folds of men usually result in a lower mean f0 than that of women [17], one variable for which mixed and at times seemingly conflicting gender associations exist is creaky voice.

1.3. Creaky Voice

Creaky voice, or creak, is a voice quality characterised by perceptual roughness or gravelyness ([18, 19]) and associated with low and irregular f0 ([18, 20, 21]). In recent years, creaky voice has been the object of unfavourable commentary in popular culture, often linking it to the speech of young women [22] and sometimes to the speech of gay men ([23, 24]). It might be assumed that this negative appraisal of creaky male voices may relate to its stereotype as feminine, and thereby its deviation from what is licensed by hegemonic masculinity.

However, scholarly investigations suggest a more complex relationship between creak and gender. Whereas early research more often linked creak to English-speaking *male* voices than female, particularly in the United Kingdom ([25, 26]), but also in Australia [27], subsequent studies in the North American context have identified and explored an association between creak and *women's* speech ([19, 28, 29]). In the AusE context, studies have demonstrated that the use of creaky voice *does* vary by gender, but that the precise nature of this variation is influenced by other facets of speaker identities, such as ethnic and language background [30], and age and race [31]. Despite this wealth of gender-based research and the pop cultural association with gay men's speech, links between creaky voice and SO have less frequently been the object of empirical investigation. Notably though, in American English speakers from California, [32] noted a connection between creaky male voices and being perceived as gay, while [33] discussed the role of creak (alongside falsetto) in facilitating a gay male speaker's adoption of a diva persona, which in turn contributed to the construal and communication of his gay identity.

One might consider that this all results in an apparent paradox. On the one hand, one might predict that speakers who ascribe to the edicts of hegemonic masculinity would avoid creak, given its popular links to feminine and gay male speech styles. On the other hand, such speakers might also be predicted to employ lower f0, given its links to being perceived as masculine ([16, 17]). This is despite low f0 facilitating creaky voice and being one of its acoustic correlates ([18, 19, 20]). However, we note that in [33] the use of creak with falsetto expanded the speaker's f0 range, which is argued to have indexed expressiveness and the speaker's diva persona.

1.4. Topic-Based Style Shifting

Lastly, topic-based style shifting refers to (usually intraspeaker) linguistic variation according to referential topic. Evidence exists that gay speakers vary their speech when moving between gay and other topics [34], but that not all gay speakers vary their speech in the same way [35]. Additional variables may impact the precise nature of speakers' style shifting. For example, the speech of those whose gender presentation is more typical may vary differently from those whose gender presentation is less so.

2. Aims

2.1. Predictions

The research presented here forms part of a larger project which aims to help close the gap in scholarly understanding of variation in AusE according to male SO. This study focused specifically on how male speakers' relationships with (hegemonic) masculinity might affect variation in the use of creaky voice. We also examined the role played by topic-based style shifting. Accordingly, we made the following three predictions:

1. Creak will be more prevalent in the speech of gay and other sexual minority men than that of straight men.

This prediction is motivated by findings of earlier studies on creak ([32, 33]). We note the complexity of gendered associations with creaky voice, as well as the apparent contradiction of creak's associations with, on the one hand, feminine and gay male voices and, on the other, low f0.

2. Men who value traditional male gender roles more highly will produce less creak than those who value traditional male gender roles less highly.

This prediction is based on the association that exists in pop culture, and in some of the scholarly literature, between creak and the voices of women and of gay men.

3. Gay men will produce more creak when speaking on an LGBTQ topic than on a neutral topic, but this will be modulated by their acquiescence to hegemonic masculinity.

This was motivated by [34] and [35]. No prediction was made regarding the straight male speakers.

3. Methods

3.1. Participants

Participants were 77 adult male speakers of AusE aged between 18 and 52 years (mean = 30.4), who reported no diagnosed speech or hearing problems and had received all primary and secondary schooling in Australia. 27 were gay, 38 were straight, and 12 reported another SO. This initial sample was later reduced to 71 following outlier removal (see below).

3.2. Tasks

Each participant took part in a single, in-person data collection session conducted by the first author in a quiet location which was recorded using a Zoom H6 Portable Recorder at a sampling rate of 44.1 kHz.

Participants engaged in three tasks: a picture description task, a video recount task, and a conversation task. Data analysed in this paper are from the picture description task, which involved the speaker describing nine different photographic images. To examine potential topic-based style shifting, the last three images were deliberately selected to depict (non-sexual) LGBTQ-themed situations. The remaining six were neutral with regard to SO and LGBTQ subjects.

Each participant completed a demographic questionnaire as well as one or more additional surveys. Participants were asked for their SO and were given the option of selecting "gay", "straight" or an open-ended field for other orientations. Other demographic information collected included the speaker's age, his ethnocultural heritage and the area(s) in which he had grown up and attended school; only the first of these is discussed further in this paper.

The additional surveys included the Male Roles Attitudes Scale (MRAS) [36], an instrument which is designed to measure an individual's concordance with statements pertaining to traditional male gender roles, and which has been deployed in previous sociophonetic studies [37]. The MRAS elicits responses to eight statements on a four-point Likert scale (1 = strongly disagree, 4 = strongly agree). Example statements are: "A man always deserves the respect of his wife and children", and "It bothers me when a man acts like a woman".

Participants reporting their SO as "gay" or "other" also responded to a series of surveys designed to measure their relationships with homosexual masculinity, though responses to these are beyond the scope of this paper.

3.3. Data Analysis

3.3.1. Demographic and Survey Data Preparation

Mean ratings across the eight MRAS items were calculated, and results z-scored (higher MRAS = more traditional male gender role values). SO was reduced to a two-way distinction between men who were straight and those belonging to the sexual minority, i.e. those who could be described as gay or otherwise not straight. Age was also simplified to create younger (≤ 30 years; N = 45) and older (> 30 years; N = 32) age brackets.

3.3.2. Speech Data Preparation

Orthographic transcriptions of the speech data were automatically generated [38] and hand corrected. These were used for automatic phonemic segmentation and labelling [39]. Macreaper [40] was used to calculate f0 estimates (in Hz) at 10 ms intervals. The union of two methods used to identify creaky voice (hereafter "Union method" [21]) was then used to determine whether or not each f0 estimate coincided with a period of creak: these two methods were the antimode (AM) method [40] and the creak detector (CD) algorithm [41]. If one or both of these methods determined that a particular time stamp was creaky, it was labelled as such. The Union method has been identified as an accurate automatic method of determining whether phonation is creaky [21]. It is most accurate when applied to sonorants only [21], and for this reason only sonorants were examined here.

The AM method ([40]) relies on low f0 to identify creak. f0 distributions are identified for each speaker. These are generally bimodal in nature with the higher peak indicating the mode f0 during modal voice, and the lower peak indicating the mode f0 during creak. The trough between these peaks is the antimode, with all 10 ms intervals with an f0 lower than the antimode labelled as creaky.

In contrast, the CD algorithm combines multiple metrics known to correlate with creak, namely H2-H1, residual peak prominence, power peak parameters, and inter-pulse similarity and inter-frame periodicity metrics [41]. A probability of creak presence is calculated for each 10 ms interval, which is converted to a binary creak decision; this decision depends on what probability threshold the algorithm is set to use. A threshold of 0.05 was used here, as has been previously used for male AusE speakers ([21, 30]).

The Union method output was used to calculate the percentage of each speaker's f0 estimates that were creaky. 6 speakers were shown to be outliers, with over 40% of their sonorant f0 estimates being creaky. Data from these outliers were excluded, resulting in a final sample of 71 speakers (straight= 36, gay= 24, other= 11; ≤ 30= 42, >30= 29).

3.3.3. Statistical Modelling

To examine which variables would predict a higher probability of creak, a logistic mixed effects regression model was constructed using the *lme4* package in *R* [42]. The presence or absence of creak at each 10 ms interval was the binary outcome variable (*creak*). Fixed factors were: *MRAS,* a continuous variable (z-scored); *SO* (*gay and other*/*straight* with reference *gay and other*); age (*≤30*/*>30* with reference *≤30*); *topic* (*neutral*/*LGBTQ* with reference *LGBTQ*). Initially, these fixed effects were included in four 3-way interactions. A random intercept was included for *speaker,* with a random slope for *topic*.

The most parsimonious model was identified by a process of model reduction, removing at each step the highest nonsignificant term with the greatest *p*-value. At each stage the resulting model was compared to the preceding one using the *anova* function to test whether the more complex model had been a significantly better fit. The final model contained a single 3-way interaction between *SO*, *MRAS* and *age*, while *topic*, having lacked significance, had been removed from the fixed effects structure entirely. The structure of this final model was:

$$creak \sim MRAS*age*SO + (1 + topic \mid speaker)$$

4. Results

There was a significant interaction between *MRAS*, *age*, and *SO* (see Table 1 for the model summary). Figure 1 shows that, amongst younger *≤ 30* speakers, *gay and other* speakers had a higher probability of producing creak than *straight* speakers. In the *≤ 30* group a higher *MRAS* score resulted in a higher probability of creak: this was true for speakers of both SOs.

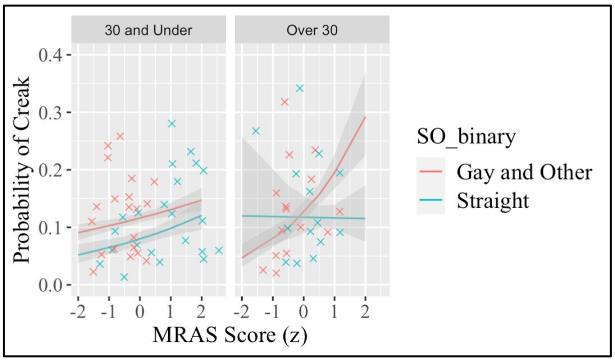


Figure 1: *Effect of SO, MRAS score and age on probability of creak*

In contrast, in the *>30* group, higher *MRAS* did not lead to a greater probability of creak for the *straight* men. However, for *gay and other* men, increasing *MRAS* led to an even greater increase in creak probability than for their *≤30* counterparts. In the older *>30* group, the model predicted a higher probability of creak for the *gay and other* speakers than the *straight* speakers when *MRAS* was higher. This was not the case, however, when *MRAS* was low.

Table 1. *Fixed Effects of GLMER*

Term	Est	Std Err	Stat	p-value
Intercept	-2.03	0.038	54.03	< 0.001***
MRAS	0.14	0.037	3.76	< 0.001 ***
age>30	0.08	0.059	1.39	0.164
SOstr	-0.42	0.064	-6.49	< 0.001 ***
MRAS:age>30	0.39	0.092	4.31	< 0.001 ***
MRAS:SOstr	0.09	0.054	1.72	0.086
Age>30:SOstr	0.35	0.085	4.15	< 0.001 ***
MRAS:age>30:SOstr	-0.64	0.102	-6.23	< 0.001 ***

5. Discussion

The results were partially consistent with our first prediction: among the younger age group, gay and other sexual minority men employed greater proportions of creak than straight men, with this also being the case among older gay men with higher MRAS scores. This would align with a characterisation of creaky voice as a feminine phonetic variant that these gay and other men, having already been marginalised from the masculine hegemony as a result of their SO, have less motivation to avoid than straight men.

However, contrary to our second prediction, among all groups except the older straight men, an increase in MRAS led to an *increased* probability of creak. This means that younger men who more highly value traditional male gender roles employ *more* creak than those who value it less highly, regardless of sexual orientation; this is also predicted for older gay and other men. This observation is clearly not reconcilable with creak carrying an exclusively feminine social meaning. Therefore, we might entertain a situation wherein creak's present feminine sociolinguistic connections emerged following women's adoption of what had thitherto been a masculine speech resource, as evidenced by findings such as [27], and that creak may retain at least some of its erstwhile masculine sociolinguistic meanings.

Additionally, in seeking to explain this apparent inconsistency, we need to revisit the acoustic characteristics of creaky voice: specifically, that it is frequently (though not exclusively) connected to low f0. It may be that the increasing probability of creak among the younger men as their MRAS scores more closely conform to traditional male gender roles could be explained by their attempting to access a particularly low f0. Should this be the case, we might theorise that this is motivated by their desire to access masculine social meanings associated with low f0. The greater probability of creak for the gay and other speakers compared to the straight speakers (also among the younger men) may also result from their accessing low f0 values while speaking. However, rather than to access social meanings situated within the indexical field of low f0 itself, the aim here may be to assist in the expansion of the f0 ranges of their utterances; this would be consistent with suggestions in [33].

The patterns of creak observed in the over-30 group are substantially different, however. Importantly, no significant effect of increasing MRAS score exists for the straight men in that age bracket. We propose possible explanations here. First, older speakers might generally be expected to be more settled and comfortable in their identities, and therefore feel less need to manipulate their voices to conform to social pressures. An alternative explanation could be that the popularity of "alpha male" influencers such as Andrew and Tristan Tate, who espouse interpretations of masculinity that are very often toxic and misogynist [43], may be resulting in a preoccupation with indexing masculinity by, for example, reducing pitch, among young men who more closely share those controversial figures' social values. The fact that the target audience of these creators skews young [43] may explain why the same effect is not seen among the straight men in the older age bracket. Notably, however, we see for the gay and other sexual minority men in that older bracket an even greater effect of increasing MRAS than in the younger bracket. Here we might suggest that, having come of age in an era when being identifiable as gay (or as otherwise possessing an SO outside the heteronormative standard) could pose a very real threat to one's reputation, livelihood or even safety [44], men in this group who did believe in traditional masculine values made a particular effort to index masculinity with their voices, such as by speaking in a low pitch and consequently with more creak. Their speech styles may have remained subsequently static, notwithstanding increasing acceptance of alternative SOs over time. We note, however, the low number of speakers in this group in our sample and suggest that further research is required to confirm this idea.

Next, even more generally, when seeking to explain the apparent inconsistency, we should also consider two theoretical perspectives that could explain potential confounds. First is an intersectional perspective. If young Australian women's creak prevalence varies according to other demographic factors such as ethnocultural background [30], it is possible, if not likely, that this, or another social variable, is interacting with SO, age and MRAS for Australian *men* too.

Second, since speakers can also access many temporary stances and attitudes using suprasegmental variants like creak, it has been argued that its social meanings ought to be decoupled entirely from (directly indexing) gender [45]. For example, it has been suggested that creak might index, in some situations, negative affect and affective disengagement or, in others, being relaxed [46] and more authoritative [30]. It is certainly conceivable that our young speakers who are more influenced by hegemonic masculinity would be more motivated to project authority than others, and, potentially, to take a relaxed or disaffected stance to avoid any feminine and gay sociolinguistic meanings associated with being animated or expressive.

Lastly, the third prediction was not supported, as the probability of creaky voice was not significantly different when speakers, whether gay and other or straight, were describing the LGBTQ themed images compared with the neutrally themed images.

6. Conclusion

This study sought to illuminate how speaker SO and orientation toward masculinity influence the production of creaky voice among Australian men, and also the role played by referential topic. Results showed that, consistent with predictions, there is a higher probability of creak in the voices of younger men who are gay or otherwise belong to the sexual minority than younger men who are straight. Contrary to predictions, however, the probability of producing creak is higher among younger men who place more value in traditional male gender roles than those who place less value in it. For older speakers the relationship is more complex. Increasing levels of agreement with male gender roles results in increased probability of creak even more than it did for the younger men, but this effect occurs only among gay and other sexual minority men. Increasing agreement with male gender roles has no such effect on the probability of creak for older straight men. Further, in terms of probability of creak, a difference between older straight men and older gay and other sexual minority men exists only for speakers with high levels of agreement with male gender roles. It is apparent, therefore, that sociolinguistic variability in the use of creaky voice is particularly complex, and is likely more nuanced than can be completely explained with reference to macro social categories. Future studies might continue the examination of its role in the construction of particular indexical stances and styles, and the effect of finer-grained elements of speakers' sexual identities.

7. References

[1] Gaudio, R. P., Sounding gay: Pitch properties in the speech of gay and straight men. *American speech, 69*(1), 30-57, 1994.

[2] Pierrehumbert, J. B., Bent, T., Munson, B., Bradlow, A. R., & Bailey, J. M., The influence of sexual orientation on vowel production. *The Journal of the Acoustical Society of America, 116*(4), 1905-1908, 2004.

[3] Podesva, R. J., & Van Hofwegen, J., How conservatism and normative gender constrain variation in inland California: The case of /s/. *University of Pennsylvania Working Papers in Linguistics, 20*(2), 15, 2014.

[4] Daniele, M., Fasoli, F., Antonio, R., Sulpizio, S., & Maass, A., Gay voice: Stable marker of sexual orientation or flexible communication device?. *Archives of Sexual Behavior, 49*, 25852600, 2020.

[5] Podesva, R. J., The California vowel shift and gay identity. *American speech, 86*(1), 32-51, 2011.

[6] Baeck, H., Corthals, P., & Van Borsel, J., Pitch characteristics of homosexual males. *Journal of Voice, 25*(5), 211- 214, 2011.

[7] Kachel, S., Simpson, A. P., & Steffens, M. C., "Do I sound straight?": Acoustic correlates of actual and perceived sexual orientation and masculinity/femininity in men's speech. *Journal of Speech, Language, and Hearing Research, 61*(7), 1560-1578, 2018.

[8] Hazenberg, E., *Liminality as a lens on social meaning: A cross-variable analysis of gender in New Zealand English*. [Doctoral dissertation] Victoria University of Wellington. Wellington, New Zealand, 2017

[9] Shea, T., Gibson, A., Szakay, A., & Cox, F., Australian English speakers' attitudes to fricated coda /t/. *Australian Journal of Linguistics, 43*(1), 87-119, 2023.

[10] Szalay, T., Holik, J., Nguyen, D. D., Morandini, J., & Madill, C. J., Comparing first spectral moment of Australian English /s/ between straight and gay voices using three analysis window sizes. In *Proceedings of the 24th INTERSPEECH* (pp. 20-24). Dublin, Ireland, 2023.

[11] Connell, R.W., & Messerschmidt, J.W., Hegemonic Masculinity: Rethinking the Concept. *Gender & Society, 19*(6), 829-859, 2005.

[12] Butler, J., Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre Journal, 40*(4), 519-531, 1988.

[13] Zwicky, A. M., Two lavender issues for linguists. In Anna Livia & Kira Hall (eds.), *Queerly phrased: Language, gender, and sexuality*, 21–34. New York: Oxford University Press, 1997.

[14] Munson, B., The acoustic correlates of perceived masculinity, perceived femininity, and perceived sexual orientation. *Language and Speech, 50*(1), 125-142, 2007.

[15] Drager, K., Hardeman-Guthrie, K., Schutz, R., & Chik, I., Perceptions of style: A focus on fundamental frequency and perceived social characteristics. In L. Hall-Lew, E. Moore, & R. J. Podesva (Eds.), *Social meaning and linguistic variation: Theorizing the third wave* (pp. 176-196). Cambridge University Press: Cambridge, United Kingdom, 2021.

[16] Campbell-Kibler, K., & miles-hercules, d., Perception of gender and sexuality. In J. S. Ehrlich, M. Meyerhoff, & J. Holmes (Eds.), *The Routledge handbook of language, gender, and sexuality* (pp. 123-136). Routledge, 2021.

[17] Seikel, J. A., Drumright, D. G., & King, D. W., *Anatomy & physiology for speech, language, and hearing* (5th ed.). Cengage Learning: Clifton Park, NY, United States of America, 2016.

[18] Dallaston, K. and Docherty, G., The quantitative prevalence of creaky voice (vocal fry) in varieties of English: A systematic review of the literature. *PloS ONE, 15*(3), 2020.

[19] Wolk, L., Abdelli-Beruh, N. B., & Slavin, D., Habitual use of vocal fry in young adult female speakers. *Journal of Voice, 26(*3), 111-116, 2012.

[20] Keating, P. A., Garellek, M., & Kreiman, J., Acoustic properties of different kinds of creaky voice. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS 2019)* (pp. 2-7). Melbourne, Australia: International Phonetic Association, 2019.

[21] White, H., Penney, J., Gibson, A., Szakay, A., & Cox, F., Evaluating automatic creaky voice detection methods. *The Journal of the Acoustical Society of America, 152*(3), 1476-1486, 2022.

[22] Swannell, C., Please stop frying your vocals, ladies. *The Medical Republic*. Retrieved from https://www.medicalrepublic.com.au/please-stopfryingyourvocals-ladies/18024 , 2023, September 26.

[23] Shariatmadari, D., Do you sound gay? What our voices tell us – and what they don't. *The Guardian*. Retrieved from https://www.theguardian.com/commentisfree/2015/jul/17/campthe-voice-gay-rights, 2015, July 17.

[24] Cushing, S., *Be you, voice and all* [Video]. TikTok. Retrieved from https://www.tiktok.com/@sam.cushing/video/6825796618541960453, 2020, December 5.

[25] Henton, C. and Bladon, A., Creak as a sociophonetic marker. In Hyman, L. M. and Li, C. N. (Eds), *Language, speech, and mind: Studies in honour of Victoria A. Fromkin* (pp. 3–29). Routledge, 1998.

[26] Stuart-Smith, J., Glasgow: Accent and voice quality. In Foulkes, P. and Docherty, G. (Eds) *Urban voices: Accent studies in the British Isles*, (pp. 203– 222). Arnold, 1999.

[27] Pittam, J., Listeners' evaluations of voice quality in Australian English speakers. *Language and Speech, 30*(2), 99113, 1987.

[28] Podesva, R. J., & Callier, P., Voice quality and identity. *Annual review of applied Linguistics, 35*, 173-194, 2015.

[29] Podesva, R. J., & Szakay, A., Gender differences in the acoustic realization of creaky voice: Evidence from conversational data collected in Northern California. *Journal of the Acoustical Society of America, 134(5_Supp)*, 4238-4238, 2013.

[30] White, H., *Creaky Voice in Australian English*. [PhD Thesis] Macquarie University. Sydney, Australia, 2023.

[31] Loakes, D., & Gregory, A., Voice quality in Australian English. *JASA Express Letters, 2*(8), 2022.

[32] Zimman, L., Hegemonic masculinity and the variability of gay-sounding speech: The perceived sexuality of transgender men. *Journal of Language and Sexuality, 2*(1), 1-39, 2013.

[33] Podesva, R. J., Phonation type as a stylistic variable: The use of falsetto in constructing a persona 1. *Journal of Sociolinguistics, 11*(4), 478-504, 2007.

[34] Levon, E., Teasing apart to bring together: Gender and sexuality in variationist research. *American Speech, 86*(1), 69-84, 2011.

[35] Dickson, V., & Turner, Y., Pulling out all the stops: Referee design and phonetic correlates of gay men's English. *Lifespans and Styles, 1*, 3-11, 2015.

[36] Pleck, J. H., Sonenstein, F. L., & Ku, L. C., Attitudes toward male roles among adolescent males: A discriminant validity analysis. *Sex Roles, 30*(7), 481-501, 1994.

[37] Levon, E., Categories, stereotypes, and the linguistic perception of sexuality. *Language in Society, 43*(5), 539-566, 2014.

[38] IBM Corporation, IBM Watson Speech to Text. Retrieved from https://www.ibm.com/watson , 2023.

[39] Schiel, F., Automatic Phonetic Transcription of NonPrompted Speech. In *Proceedings. of the International Congress of the Phonetic Sciences* (pp. 607-610), San Francisco, USA, 1999.

[40] Dallaston, K., & Docherty, G., Estimating the prevalence of creaky voice: A fundamental frequency-based approach. In *Proceedings of the 19th International Congress of Phonetic Sciences. Australasian Speech Science and Technology Association Inc* (pp. 532-536). Melbourne, Australia, 2019, August.

[41] Drugman, T., Alku, P., Alwan, A., & Yegnanarayana, B., Glottal source processing: From analysis to applications. *Computer Speech & Language, 28*(5), 11171138, 2014.

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

[43] Wescott, S., & Roberts, S., Andrew Tate's extreme views about women are infiltrating Australian schools. We need a zero-tolerance response. *The Conversation*. Retrieved from https://theconversation.com/andrew-tates-extremeviews-aboutwomen-are-infiltrating-australian-schools-we-need-a-zerotoleranceresponse-229603 , 2014, May 10.

[44] Tomsen, S., Masculinity and homophobic violence in Australia's recent past. *Sexuality & Culture, 21*, 813-829, 2017.

[45] Becker, K., & Zimman, L., Beyond binary gender: creaky voice, gender, and the variationist enterprise. *Language Variation and Change, 34*(2), 215-238, 2022.

[46] Ligon, C., Rountrey, C., Vaidya Rank, N., Hull, M., and Khidr, A., Perceived desirability of vocal fry among female speech communication disorders graduate students. *Journal of Voice, 33*(5). 805.e21–805.e35, 2019.

FSEEL: A Platform for Listening-Based Evaluation of Forensic Speech Enhancement

Vincent Aubanel, Helen Fraser

Research Hub for Language in Forensic Evidence, University of Melbourne
vincent.aubanel@unimelb.edu.au, helen.fraser@unimelb.edu.au

Abstract

This paper introduces FSEEL, a platform for evidence-based evaluation of Forensic Speech Enhancement (FSE). Current practices in handling poor-quality audio evidence in legal settings are often inadequate, leading to potential miscarriages of justice. FSEEL addresses these issues by providing an integrated interface for comparing and assessing enhanced speech samples. The platform emphasises the need for quantifiable data, reproducible methods, and explicit processing chains to bridge the gap between audio engineering expertise and legal requirements. By combining FSE with forensic transcription in a single interdisciplinary approach, FSEEL aims to improve the reliability of enhanced audio evidence in court.

Index Terms: forensic speech, speech enhancement, speech intelligibility

1. Background

Poor quality audio is frequently presented as crucial evidence in court. However, current legal practices for handling such evidence are problematic, often leading to serious injustices [1]. The legal system treats audio enhancement and transcription as separate processes, both of which are frequently mishandled [2]. This situation persists largely because legal professionals – lawyers, judges, and jury members – evaluate these elements without specialised knowledge in speech perception. A key issue is the lack of recognition of the complex interaction between contextual knowledge, audio enhancement, and transcription. The legal approach, based on precedent rather than scientific principles, fails to account for these crucial relationships. Australian linguists have been actively working to reform these laws, aiming to ensure all forensic audio evidence is provided with demonstrably reliable enhancement and transcription before trial [3]. However, implementing such reforms is challenging as current language and speech sciences cannot provide an off-the-shelf solution for forensic speech enhancement in legal contexts [4]. To address these challenges, the Research Hub for Language in Forensic Evidence at the University of Melbourne proposes to combine Forensic Speech Enhancement (FSE) and Forensic Transcription (FT) into a single interdisciplinary project. This approach uses a listening-based methodology, whereby rigorously tested listeners follow an evidence-based method to produce demonstrably reliable transcripts to accompany responsibly processed audio. This solution represents a significant shift from current practices, and requires collaboration between audio engineering experts, linguists, and legal professionals. It aims to create a scientifically grounded approach to forensic audio evidence, addressing the current disconnects between enhancement and transcription in particular. This paper focuses on the enhancement branch of this project, exploring how an integrated, evidence-based approach to FSE can contribute to more reliable outcomes in legal proceedings.

2. The problem of FSE

2.1. Enhancement in Speech Research

In speech enhancement research, continued efforts in the last decades have achieved impressive results [5, 6], with recent advances using Machine Learning techniques being able to virtually removing mild levels of noise from a speech signal [7, 8, 9, 10]. On the main, speech enhancement research has been driven by common real-life applications, such as removing audible artefacts, attenuating background noise, or separating competing sources. That is, most of the efforts have be aimed at improving speech *quality*, typically occurring with positive Signal-to-Noise Ratios (SNRs). A subset of that research has focused on improving speech *intelligibility*, tackling more adverse conditions (i.e. worse SNRs), with there too positive results, albeit using more controlled speech material [11, 12, 13].

Recent work using causal diffusion networks [14, 15] make it possible to consider new types of distortion outside the traditional additive noise and reverberation scenarios, as well as ability to operate at worse SNRs, making it suitable to tackle more forensic-like situations. While this is a promising approach, we note that suggested solutions to overcome current limitations of these approaches involve adding supplementary information originating from the target material – which is by definition not an option when handling a forensic speech recording. More generally, the question of the evaluation has become central to modern enhancing techniques [16] – and is especially relevant in the forensic context.

2.2. Forensic Speech is special

As introduced in [4], forensic speech departs in significant ways from the speech commonly considered in general speech enhancement research.

Recording conditions are generally far worse than the ones used for typical speech enhancement research programmes. This is because forensic speech is not a research object, with carefully designed degradations. The criterion that makes a recording a sample of interest is that it is deemed to contain incriminating evidence. The acoustic characteristics are therefore largely independent of any theoretical research question, and the levels of degradation can often exceed the level of recoverability.

The difference in degradation is not only quantitative, but also qualitative. Indeed, the type of degradation typically encountered in forensic settings are extremely variable, which make it difficult to propose a unified enhancing approach, but rather calls for ad-hoc solutions. In fact, this dependence on essentially non-replicable procedures exerts pressures on FSE practice to escape scientific approach, by preventing the collection of large number of examples of similar quality, in order to develop efficient research methods.

The lack of ground truth is both a technical limitation, as the absence of a reference signal or transcript prevents the use of existing scoring methods, and also a conceptual one as it makes harder to specify the objective, and formulate successful enhancing strategies.

Another specificity of forensic speech is that it entertains a much broader context than traditionally-considered speech material. That is, intelligibility can sometimes only be established on the basis of information external to the recording itself – a fact that is usually ignored, or conveniently controlled for, in traditional speech enhancement frameworks. Linked to this point is the fact that while the target "enhanced" speech form is often assumed to be intelligible, with the challenge of enhancing being to overcome externally applied distortions, large part of forensic speech material has very low *intrinsic* intelligibility owing to talker factors, such as poor articulation, interrupted speech or foreign accent, to mention just a few examples. This is important to keep in mind when defining the objective of any enhancing method.

Finally, an essential difference between speech enhancement and FSE relates to their very purpose: while in speech enhancement aim is to model human speech perception, pushing knowledge boundaries at every iterations, the purpose of FSE is strikingly different, as it seeks to support reliable transcription of a specific potentially ambiguous speech production, to enable the court to understand what was said, in the most unambiguous way possible. A corollary of this is that while removing background noise could certainly be helpful in the general situation, and help retrieving what was said by increasing SNR, noise reduction should not constitute an objective of FSE by any measure: firstly noise reduction is commonly found to introduce artefacts [17], usually resulting in a net intelligibility *decrease*, but in some forensic scenarios, the noise itself might contain evidence or contextual information. Therefore, focussing too strongly on noise reduction might disserve the overall purpose of forensic speech enhancement. Further, it has been found that enhancement treatments that loose sight of the forensic purpose result in intelligibility *decrease*: speech perceived to have been enhanced receive increased credibility, and while fewer words are identified experimentally in this condition, it is the enhanced form that is preferred by listeners – with obvious negative consequences in real court situations [18, 1].

The purpose of FSE might therefore include relying on strong parsimony measures in terms of performing change on the speech signal, and should perhaps by guided by an overall aim of intelligibility increase, while at the same time minimising affecting the samples' identity [19].

2.3. Current practices in Forensic Speech Enhancement

In addressing the specific challenges of forensic speech defined above, practitioners have employed what could be called a "toolbox approach" to speech enhancement, that is the application of a set of mitigation strategies tailored to a set of challenges [20]. This has been served by plugins in commercially available software solutions (e.g. iZotope RX, ACON Digital Acoustica, WAVES Audio) or specific forensic systems (CEDAR Audio, Salient Sciences, Cube-Tech). Proposals to structure these strategies and make them more reproducible have been made, for example by suggesting sequential application of processing [21]. In practice however, and without the explicit formulation of objective intelligibility-increase goals informed by evidence-based research methods, the evaluation of speech enhancement is often left to rely on unspecified subjective listening by otherwise highly-talented practitioners, that is, with skills mismatched to that of intelligibility evaluation.

A number of Artificial Intelligence (AI)-inspired tools are also becoming increasingly commercially available, either integrated in the software suites mentioned above or as standalone online resources (e.g., https://hance.ai, https://www.lalal.ai/voice-cleaner/, https://dolby.io/products/enhance/). In contrast to the "toolbox approach", these solutions offer very limited set of control parameters – if any – to enhance speech samples. While the result can sometimes be spectacular, these approaches can be problematic, but from an opposite reason: the practitioner is deprived of the knowledge of the enhancement deployed and cannot justify or explain the modifications that the speech sample has undergone.

The current resulting situation is that enhanced samples which end up in court can be of surprising low quality. This also arises as a consequence oflegal procedures whereby practitioners are required to demonstrate adherence to specific submission criteria, as opposed to scientifically defined criteria for listenability or intelligibility increase [4] .

2.4. The requirements for Forensic Speech Enhancement

There is a need to establish proper evaluation methods for forensic speech enhancement output. On one hand, mature methodology exists in speech research to evaluate enhancement output. On the other hand, expert and ad-hoc approaches are by nature more difficult to capture in replicable and quantitative ways. The goal of the current work is therefore, alongside other work concerned with improving forensic transcription, to establish quantifiable evidence-based method for FSE.

3. FSEEL: Description of the platform

3.1. Definition

Forensic Speech Enhancement for Expert Listening (FSEEL) is a platform aiming at centralising speech enhancement procedures in an integrated interface, enabling streamlined and explicit comparison of processed samples, with a view to fostering a reproducible evidence-based approach. Its purposes and features are described in detail below, but before it is perhaps useful to highlight what it is *not*, to emphasise its functionalities in contrast to existing tools. First, FSEEL is not a *signal* editor, which would allow to explore interactively (i.e., zooming in, selecting portions of the signal etc.). Here, while some degree of flexibility is allowed to interact with samples, the focus is to limit distractive interaction with the signal, and instead to provide visual information, to minimise uncontrolled subjective interpretation of enhancement. Second, FSEEL is not a *transcription* platform. It should rather seen as a preparatory platform, which aim is to process a sample that will undergo a transcription in a later stage. In fact, in has been designed to integrate with Soundscribe, a dedicated platform for transcription of forensic speech (in development in our group). . Finally, FSEEL is not a *modification* platform. Here, all enhancement processes are performed offline. One reason is practical: given the variety of speech enhancement that the platform is aiming to include, it can be difficult to enable real-time access to specific resources. The main reason however is conceptual: we wanted to keep the focus on an evidence-based workflow in contrast to promoting iterative trial-and-error modification loops.

3.2. Purposes

In its early stage, it is as much an exploratory as a conceptual tool: while it allows users to simulate and visualise the effect of enhancement in very concrete terms – as it enables to listen to the different versions at all stages of processing and read the associated analysis data – it also fosters thinking about the processing chain as a whole, and develop an understanding of the various processes and their impact at different stages of processing, specifically in terms of relative intelligibility and quality gains.

One of the main goal of the platform is to provide a unifying interface for sometimes disparate sources of sound processing. For example, a practitioner may have some preference for specific processing with third-party application, and may wish to include this processing alongside standardised formatting. The current solution aims to provide the integration of third-party enhancement as widely as possible, through the modular specification of processing. It can interact natively with most popular sound processing libraries and tools (e.g. *sox*, *ffmpeg*) and currently supports commercially available plugins in the VST3 and AU format.

Extending its organisation as a processing chain, the platform is also destined to be used "in reverse", that is, as a way to generate degraded speech in forensically meaningful ways. This is an important dimension of our project, and will build on earlier localised initiatives (e.g. [22] and [14])

Finally, we designed the interface to be as explicit as possible with regard to changes made to the speech signal. The user should have a clear idea of exactly the type of processing the sample goes through, and is aided with relevant quantified numerical information associated with each step. This is to help constraining and guiding the expectations of enhancement, and help evaluating their effect at any given step.

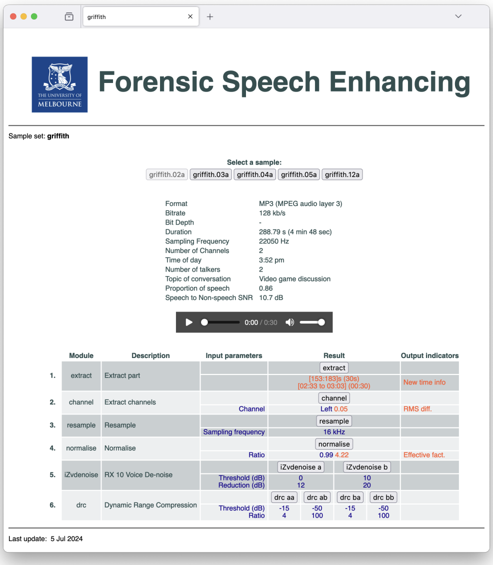
3.3. Features

Figure 1 shows a screenshot of the platform, illustrating current features for comparing and exploring various speech enhancements.

The platform allows to handle a set of samples, which is a typical situation, e.g., several extracts taken from different point in time of a lengthy recording. All processing steps are identical for each sample, but the value of parameters can be specified individually for each sample – e.g., it can be desirable to specify different level adjustment strategies for different portions of a recording.

As a first source of visual information, a detailed list of information is displayed relating to the sample, in its original (unprocessed) form, including technical information such as overall duration, file format, file size or sampling rate. Other metadata related to the content of the recording are displayed here as well, such as the number of talkers or the topic of conversation. If a transcription is available, a range of metrics can be computed automatically and displayed here, such as proportion of speech, or speech-to-nonspeech signal-to-noise ratio. Importantly though, this section is meant to be curated in a careful way, in order to control contextual information that is available to the user at this stage of the handling of forensic audio. The architecture is flexible and allows to implement specific context-management strategies (e.g. [23]).

Below the list of sample-specific information, an audio player is inserted which allows to playback the sample. The list of processing steps is displayed as a table where each row is a module. This emulates the processing chain that is commonly found in Digital Audio Workstations, and indicates strict sequential processing. For each module, its name and a short description are given in the columns with those respective names, and the resulting effect can be listened to by clicking on the corresponding button in the *Result* column, which loads the corresponding sample in the audio player.



Forensic Speech Enhancing

Sample set: **griffith**

Select a sample:

Format	MP3 (MPEG audio layer 3)
Bitrate	128 kb/s
Bit Depth	-
Duration	288.79 s (4 min 48 sec)
Sampling Frequency	22050 Hz
Number of Channels	2
Time of day	3:52 pm
Number of talkers	2
Topic of conversation	Video game discussion
Proportion of speech	0.86
Speech to Non-speech SNR	10.7 dB

	Module	Description	Input parameters	Result	Output indicators
1.	extract	Extract part		extract [153:183]s (30s) [02:33 to 03:03] (00:30)	New time info
2.	channel	Extract channels	Channel	channel Left 0.05	RMS diff.
3.	resample	Resample	Sampling frequency	resample 16 kHz	
4.	normalise	Normalise	Ratio	normalise 0.99 4.22	Effective fact.
5.	iZvdenoise	RX 10 Voice De-noise	Threshold (dB) Reduction (dB)	iZvdenoise a: 0, 12 iZvdenoise b: 10, 20	
6.	drc	Dynamic Range Compression	Threshold (dB) Ratio	drc aa: -15, 4 drc ab: -50, 100 drc ba: -15, 4 drc bb: -50, 100	

Last update: 5 Jul 2024

Figure 1: *A screenshot of the platform. See text for description.*

The list of modules is thought to be flexible, to allow the inclusion of processing from a wide range of sources, from custom scripts, commercial third-party VST3 plugins to offline manual editing. Additionally, any module can marked to be excluded from being displayed, while still being included in the processing chain.

Adjusting the parameters of each module and visualising their effect is an essential feature of FSEEL in supporting expert listening, and they are colour-coded, in blue and orange respectively. Only a selection of input parameters and output indicators are displayed with their units in their respective column, and their value is displayed in the *Result* column for quick visual evaluation.

Forking is one of the main feature of the platform. It allows to compare the differential effect of a module by varying a specific parameter, or a combination of parameters, and propagate the effect down the chain, therefore fostering replicable and tractable enhancement, while maintaining full flexibility. Forking is implemented visually simply by splitting cells below where the forking is specified, and each branch is visually identified in the button name by appending a letter to the label of the playback button. For example, varying three parameters for a given module will be displayed by inserting suffixes 'a', 'b' and 'c' to each branch down the module list respectively.

The implementation of the platform currently consists of a processing backend currently written in MATLAB, as many speech-related applications are available in that language, and a frontend consisting of a html webpage, for wide compatibility. Other languages can be envisaged to optimise future functionality (e.g., interactive processing) and interface with external speech processing libraries.

3.4. Preliminary outline of our first project

We sketch here a typical use of FSEEL for selecting samples to include in a future perception experiment, demonstrating an evidence-based approach to the evaluation of FSE. We employed the Griffith Corpus of Spoken Australian English (GC-SAusE) Collection [24], an Open Access resource of conversational speech, which satisfies a number of criteria representative of forensic scenarios: multiple speakers, varying quality of recording, diverse range of background noises and uncontrolled topic of conversation. Importantly, Conversation Analysis (CA)-style transcripts are available, offering a detailed account of the content of the conversation, which we can use as a reference transcription, upon checking its accuracy.

As seen on Figure 1, a few samples from the Griffith corpus are loaded and for the selected sample, basic technical as well as contextual information are displayed, allowing to quickly acquire an understanding of the auditory sample and guide, in evidence-base ways, the expectations of enhancement processes.

This example shows six stages of processing, sequentially ordered from top to bottom. Forking is illustrated in both steps 5 and 6, and the user can intuitively navigate the resulting output at each step and each branch of this processing tree. Selected parameters (in blue) and their resulting measured effect on the sample (in orange) allow for a rapid, objective evaluation at a glance, even before listening, of the relevant parameter being manipulated at each step. We expect that this explicit and exhaustive approach to the enhancing chain will contribute to minimise subjective evaluation effects that are bound to happen when relying exclusively on the auditory modality. In particular, it will assist in focusing attention to specific and explicit signal manipulations, while allowing the user to relate their individual effect to the general objective of (potential) global net intelligibility increase. This approach could also help by reducing the need for replay which, while having been found to increase intelligibility, can also increase overconfidence [25].

Following critical listening, promising samples are selected to form a stimuli base for a listening experiment, to be deployed through Soundscribe (see above). Promising samples in the current example involve identifying regions for which a differential in intelligibility is identified, i.e., when a difference in intelligibility is detected between different steps and branches, which could be attributed to a specific enhancing process.

We think that this explicit approach to enhancing can be of interest to the various and diverse communities involved in forensic enhancing, by illustrating to each the various exponents of speech perception. To the audio engineering community it can suggest a replicable methodology and can emphasise the interaction of enhancing with transcription and context. To legal parties it may clarify that enhancing is not the result of a single on/off button, and does not always result in net intelligibility gains.

3.5. Potential later developments

In its current state, the purpose of the platform is to attract interest from various communities involved and initiate a discussion to solve issues around forensic speech enhancement in an end-to-end approach. Future developments will therefore hopefully be shaped by those interactions. We can however already forsee various improvements that can be done, some of which we detail hereafter.

There is an increase of User Generated Recordings (UGRs, [26]) in forensic scenarios which often come with a video modality. Including capacity to toggle video display could be useful in assisting critical listening, provided context is managed appropriately [23]. Efficient synchronisation of different sources in that context, or of externally enhanced versions of samples would seems a needed feature, which could be largely automatised with existing libraries ([27], [28]).

Finally, an important function of the platform will be to generate an exhaustive contextualised report, to be carried forward alongside the resulting enhanced audio sample. The specifications of the content and the format will have to be jointly established with various actors of the domain.

4. Discussion and further planned developments

In a more conceptual perspective, we plan to extend the functionality of the platform by incorporating to the *Output parameters* a selection of evaluation metrics for each stage of processing. In addition to the battery of traditional objective speech quality metrics such as PESQ, it would be interesting to integrate, for every module, an estimation of *expected* intelligibility or quality increase based on available experimental listener data.

Given the rapid pace of developments in Artificial Intelligence (AI) and its application to speech technology, it is not unreasonable to think that the general speech enhancement problem may be reformulated in the future in terms of a speech separation problem (see, e.g. [29, 15]), at least for favourable signal-to-noise ratios, and for commonly encountered disruptions. In this reformulation, target speech would be a stem in the mixture of sources to separate, and the aim of enhancement would be to segregate each source in order to identify potential incriminating information. The modules in the platform could as a consequence look quite different, i.e. Declicking and Denoising processed might be replaced by relative mixing of varying sources. The functionality of the platform will remain the same in providing evaluating capacities to enhancing processes. In fact, the purpose of a platform such as FSEEL, appears to be even more crucial as the performance of speech enhancement progresses. Indeed, the type of output made in the general framework of generative AI commonly equates or exceeds naturally produced stimuli. We can therefore surmise that there will be increased level of confidence associated with such enhanced output – including when the enhanced version is misleading.

5. Conclusion

In this paper we introduced FSEEL, a platform dedicated to address the challenges posed by Forensic Speech Enhancement. By focusing on quantifiable data and reproducible methods, it seeks to establish the basis for a listening-based evaluation of forensic speech enhancement, which we propose is a necessary step for subsequent reliable transcription. Ultimately, this will permit a more robust, transparent, and just approach to handling forensic audio evidence in the legal system.

6. Acknowledgments

We thank two anonymous reviewers for their helpful comments.

7. References

[1] Fraser, H. and Kinoshita, Y., "Injustice Arising from the Unnoticed Power of Priming: How Lawyers and Even Judges can be Misled by Unreliable Transcripts of Indistinct Forensic Audio," *Criminal Law Journal*, vol. 45, no. 3, pp. 142–152, 2021.

[2] Fraser, H., "Enhancing Forensic Audio: What Works, What Doesn't, And Why," *Journal of Law and Human Dignity*, vol. 8, no. 1, pp. 85–102, 2020.

[3] ——, "Forensic Transcription: Legal and scientific perspectives," in *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications*, ser. Studi AISV, Bernardasci, C., Dipino, D., Garassino, D., Negrinelli, S., Pellegrino, E., and Schmid, S., Eds. IT: Officinaventuno, 2022, vol. 8, pp. 19–32.

[4] Fraser, H., Aubanel, V., Maher, R. C., Mawalim, C. O., Wang, X., Pocta, P., Keith, E., Chollet, G., and Pizzi, K., "Forensic speech enhancement: Towards reliable handling of poor-quality speech recordings used as evidence in criminal trials." *Journal Of The Audio Engineering Society*, 2024.

[5] Yu, D., Gong, Y., Picheny, M. A., Ramabhadran, B., Hakkani-Tür, D., Prasad, R., Zen, H., Skoglund, J., Černocký, J. H., Burget, L., and Mohamed, A., "Twenty-Five Years of Evolution in Speech and Language Processing," *IEEE Signal Processing Magazine*, vol. 40, no. 5, pp. 27–39, Jul. 2023.

[6] O'Shaughnessy, D., "Speech Enhancement—A Review of Modern Methods," *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 1, pp. 110–120, Feb. 2024.

[7] Luo, Y. and Mesgarani, N., "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[8] Schröter, H., Rosenkranz, T., Escalante-B., A. N., and Maier, A., "DeepFilterNet: Perceptually Motivated Real-Time Speech Enhancement," May 2023.

[9] Strauss, M., Pia, N., Rao, N. K. S., and Edler, B., "SEFGAN: Harvesting the Power of Normalizing Flows and GANs for Efficient High-Quality Speech Enhancement," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2023.

[10] Richter, J., Welker, S., Lemercier, J.-M., Lay, B., and Gerkmann, T., "Speech Enhancement and Dereverberation with Diffusion-based Generative Models," Jun. 2023.

[11] Cooke, M., Mayo, C., and Valentini-Botinhao, C., "Intelligibility-enhancing speech modifications: The hurricane challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.

[12] Aubanel, V. and Cooke, M., "Information-preserving temporal reallocation of speech in the presence of fluctuating maskers," in *Proc. Interspeech*, 2013, pp. 3592–3596.

[13] Cooke, M., King, S., Garnier, M., and Aubanel, V., "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, Mar. 2014.

[14] Serrà, J., Pascual, S., Pons, J., Araz, R. O., and Scaini, D., "Universal Speech Enhancement with Score-based Diffusion," Sep. 2022.

[15] Richter, J., Welker, S., Lemercier, J.-M., Lay, B., Peer, T., and Gerkmann, T., "Causal Diffusion Models for Generalized Speech Enhancement," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 780–789, 2024.

[16] de Oliveira, D., Welker, S., Richter, J., and Gerkmann, T., "The PESQetarian: On the Relevance of Goodhart's Law for Speech Enhancement," Jun. 2024.

[17] Hilkhuysen, G., Gaubitch, N., Brookes, M., and Huckvale, M., "Effects of noise suppression on intelligibility. II: An attempt to validate physical metrics," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 439–450, Jan. 2014.

[18] Fraser, H., "Don't believe your ears: 'enhancing' forensic audio can mislead juries in criminal trials," *The Conversation*, 2019.

[19] Aubanel, V., "A system to enhance the listening process," Master's thesis, University of Limerick, 2003.

[20] Scientific Working Group on Digital Evidence (SWGDE),, *Best Practices for the Enhancement of Digital Audio*. (v. 20-A-001-2.0), https://www.swgde.org/documents/draft-released-for-comment/, 2023.

[21] Zjalic, J., "A Proposed Framework for Forensic Audio Enhancement," Master's thesis, University of Colorado, 2017.

[22] Sharma, D., Hilkhuysen, G., Gaubitch, N. D., Brookes, M., and Naylor, P., "C-Qual—a validation of PESQ using degradations encountered in forensic and law enforcement audio," *Journal Of The Audio Engineering Society*, no. 8-1, Jun. 2010.

[23] Dror, I., Thompson, W., Meissner, C., Kornfield, I., Krane, D., Saks, M., and Risinger, M., "Context Management Toolbox: A Linear Sequential Unmasking (LSU) Approach for Minimizing Cognitive Bias in Forensic Decision Making," *Journal of Forensic Sciences*, vol. 60, pp. 1111–1112, Jul. 2015.

[24] Haugh, M. and Chang, W.-L. M., "Collaborative creation of spoken language corpora," in *Pragmatics and Language Learning*. T. Greer, D. Tatsuki & C. Roever (Eds.): University of Hawai'i at Mānoa: National Foreign Language Resource Center, 2013, vol. 13, pp. 133–159.

[25] Hilkhuysen, G., Lloyd, J., and Huckvale, M., "Effects of replay on the intelligibility of noisy speech," in *Proc. of 46th AES Conf.*, 2012.

[26] Miller, B. F., Robertson, F. A., and Maher, R. C., "Forensic Handling of User Generated Audio Recordings," *Journal Of The Audio Engineering Society*, no. 10515, Oct. 2021.

[27] Joren, S. and Leman, M., "PANAKO - A Scalable Acoustic Fingerprinting System Handling Time-Scale And Pitch Modification," *Proc. of 15th ISMIR*, 2014.

[28] Ellis, D. P. W., "Robust Landmark-Based Audio Fingerprinting," 2012.

[29] Pons, J., Liu, X., Pascual, S., and Serrà, J., "GASS: Generalizing Audio Source Separation with Large-scale Data," Sep. 2023.

Evaluating Transcripts of Poor-Quality Forensic Audio: Sine-Wave Speech and Forensic Audio

Eleanor Kettle, Helen Fraser

Research Hub for Language in Forensic Evidence, University of Melbourne
eleanor.kettle@student.unimelb.edu.au; helen.fraser@unimelb.edu.au

Abstract

Sine-wave speech (SWS), created by combining sine-waves that track formants, initially sounds completely unlike speech, but priming listeners with the original sentence makes the words 'pop out' dramatically. Forensic audio shows a similar pop-out effect when a transcript is given – even if the transcript is inaccurate. In order to explore some of the nuances of the priming effect, this paper reports a new SWS experiment, in which the original sentence was only suggested, not given as the 'right answer'. Results are discussed in the context of forensic transcription, asking how responses can be evaluated in the absence of 'ground truth'.

Index Terms: sine-wave speech; forensic transcription; human speech perception

1. Introduction

Forensic speech recordings provide powerful evidence in criminal trials, but the audio is often of such poor quality that the jury needs a transcript to understand it. Current Australian law allows transcripts to be provided by police investigating the case [1]. To protect the jury from being misled, lawyers and judges check the transcript against the audio. This is ineffective, as listeners can be primed to hear in line with a transcript even if it is inaccurate [2] – see 90-second video at forensictranscription.net.au.

ASSTA is part of a 'call to action' asking the law to recognize forensic transcription as a science. This raises an interesting topic for speech science: how to reliably evaluate the accuracy of a transcript of poor-quality audio *in the absence of a known 'right answer', or ground truth.* The present paper explores some relevant issues by reviewing a famous 1981 experiment studying the perception of 'sine-wave speech' [3], then adding to its results with findings from a new experiment. Experiment files are at: https://www.mrc-cbu.cam.ac.uk/personal/matt.davis/sine-wave-speech/.

2. Sine-wave speech

2.1. What is sine-wave speech?

Sine-wave speech (SWS) was developed at Haskins Laboratories in the 1970s to provide non-speech files for use in experiments investigating human speech perception [4]. It is produced by combining two or three sine-waves that track the formants of an utterance. On first hearing, most people find it completely unlike speech: a series of beeps and whistles suggestive of some electronic source. However, when they are primed with the original sentence, most listeners find the words 'pop out' from the meaningless noise. This effect was first published in the original 1981 study by Remez et al [3].

2.2. The 1981 study

The 1981 study used a recording of the sentence: 'Where were you a year ago', to create seven SWS versions (audio not available): one with sine-waves tracking all three formants; three with two sine-waves tracking formants 1 and 2, 1 and 3, 2 and 3; and three with single sine-waves.

Each of these seven SWS versions were played under three conditions: Condition 1: participants given no information; Condition 2: participants told it was computer speech and asked if they could discern the content; Condition 3: participants told it was 'Where were you a year ago' produced by a computer, and asked how clear they found it.

One sentence, seven versions, and three conditions gives 21 experiments, deployed to 21 different groups of 18 undergraduate psychology students. Only versions with sine-waves tracking (a) F1-3, or (b) F1-2, gave useful results, so we focus on those here, for a total of 36 participants.

2.3. Results of the 1981 study

In Condition 1 (no information), only two of 36 participants, both listening to the 3-wave version, heard the full sentence. Nine (25%) thought it could be human or computer speech with various distortions. The others heard 'science fiction noise', 'radio interference' or other non-speech noises. In Condition 2 (speech suggested), nine (25%) heard the full sentence; 17 (47%) correctly heard varying numbers of syllables (average around 2.5 out of 7 total syllables); and ten (28%) still heard no words. In Condition 3 (sentence suggested), most listeners heard most words, with higher scores for the 3-wave version.

To the researchers, this showed that words could be heard in the absence of speech cues previously thought to be necessary to retrieve the vocal tract configuration that had produced the sounds. 'The results of this study cannot be explained within the framework of existing theories of speech perception, for the tones contained none of the elemental acoustic cues typically held to underlie speech perception'. (p.949)

2.4. Top-down vs bottom up

Since then, these results have generally been interpreted as strong support for the powerful role played by top-down processing in speech perception – though more sophisticated bottom-up processing, such as auditory scene analysis, has been defended by some [5].

A particularly thorough argument for the role of top-down processing was made in a 2007 paper by Davis and Johnsrude [6], reporting the 1981 results in the context of many other top-down effects: '… evidence supports highly interactive processes with top-down information flow often driving and constraining interpretation.' (p.133)

A strong role for top-down processing is also supported by research on forensic transcription. However this work indicates that *inaccurate* transcripts can prime a 'pop out' effect as easily as *accurate* transcripts can. This led us to ask some questions about the 1981 results that have not yet, to our knowledge, been addressed in the SWS literature.

2.5. Some questions not yet addressed by SWS research

It is notable that SWS researchers have focused on the dramatic increase in participants who *do* hear the original sentence after it is suggested. This trend has been particularly prominent since 2007, when Davis and colleagues produced a set of SWS sentences which have received a great deal of media attention over many years [7-9]. SWS is typically presented via multimedia demonstrations, playing the incomprehensible SWS files, then the original sentence, then the SWS file again – which is now immediately accepted as, and assumed to be, the original sentence.

However this focus on an all-or-nothing effect is something of an overstatement. This is notable even in the 1981 experiment, where two participants heard the sentence before being told the file was speech; and after the suggestion, only 25% heard the sentence precisely as suggested, while 47% heard different words, and 28% heard no words at all.

The current experiment aimed to explore these nuances, by finding out what participants hear in SWS files if they are not provided with a single 'right answer'.

3. The current experiment

Our method was similar to that of the 1981 study, but also influenced by the design of forensic experiments such as [2].

We used the five 3-wave SWS files created by Davis and colleagues (link in Section 1). These were deployed via Qualtrics, under two sequential conditions. In Condition 1, participants were told they would be listening to 'distorted speech' and asked to write down what they heard (bypassing the 1981 condition of listening with no information). Condition 2 primed them with the suggestion (given via text, not audio) that 'Some people have suggested that they hear [the original sentence]'; and asked to listen again, as often as they wished, stating (a) whether they agreed; and (b) if not, what they heard instead.

A total of 52 participants completed the experiment. Most were undergraduates or recent graduates, with 69% aged 18-25; 25% aged 26-45 and 6% over 46. Most (31 of 52, 60%) were monolingual English speakers ('L1'), while 21 (40%) had various other language backgrounds ('other'), including 7 participants who self-reported as bilingual.

Participants were randomly divided into two groups (no statistically significant differences in age, gender, language background, headphone use, etc.). Group A listened first to all five SWS files under Condition 1, and then to each file under Condition 2. Group B listened to each file, first under Condition 1 and then under Condition 2. This gave them feedback on their perception of each file before they heard the next. All participants heard the files in the same order.

Participant responses were scored similarly to [10], with 3 = exactly like original; 2 = very close to original; 1 = some words but not close; 0 = no response. This gives 15 as the maximum possible score. Before-scores are the scores from Condition 1 (before original sentence suggested). After-scores are the scores from Condition 2 (after original sentence suggested – immediately after for Group B; at the end for Group A).

4. Results

4.1. Priming effect

Unsurprisingly, after-scores were markedly higher than before-scores (Figures 1 and 2), confirming that the suggestion had a strong priming effect, whether provided immediately or later. Notably, however, no suggestion was accepted by every participant. Indeed, some were explicitly rejected by more than half. Various factors appear to have affected the scores.

4.2. Participant effects

Before-scores varied greatly among participants (Figure 1). As in previous forensic experiments [10], this cannot be explained solely by language background. On average 'L1' before-scores were higher than 'other' (p=0.049), but some 'L1' participants did very poorly (e.g. two scored 0), while some 'other' participants did very well (1x12; 1x9; 3x7). Also as in previous studies (see also [11]), no other demographic characteristic we measured correlated with score. This suggests that the ability to decipher distorted speech with no contextual or textual suggestion reflects individual aptitude rather than specific demographic characteristics, in line with previous studies [10].

After-scores, though higher overall, were also variable (Figure 2), and also not correlated with language background or any other demographic characteristic we measured.

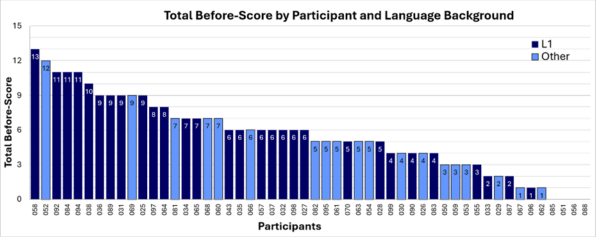


Figure 1: *Before-scores for all 52 participants.*

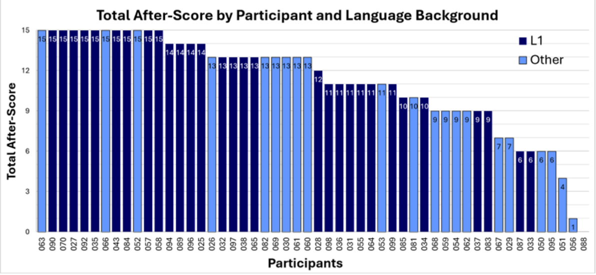


Figure 2: *After-scores for all 52 participants.*

4.3. Learning effect

Since 'other' here is a broad category with small numbers, and the effect of language background is not our present focus, all following results are reported only for the 31 'L1' participants (Group A: n=14; Group B: n=17). These smaller groups shared similar demographics to the overall participant field.

The 31 'L1' before-scores varied greatly by Group. As shown in Figure 3, for Group A (suggestion at end), highest scores were 11 and 9, most scored 6 or under, and two scored 0 (no words heard in any sentence). For Group B (suggestion after each presentation), the highest score was 13, only six scored 6 or under, and none scored 0.

It seems Group B benefitted from the suggestion after each presentation, providing formative feedback that helped them understand subsequent SWS files before the suggestion. This SWS learning effect, which is very evident to those who experience multiple SWS examples, is also reported by others [6]. However, results are not uniform, even for Group B.

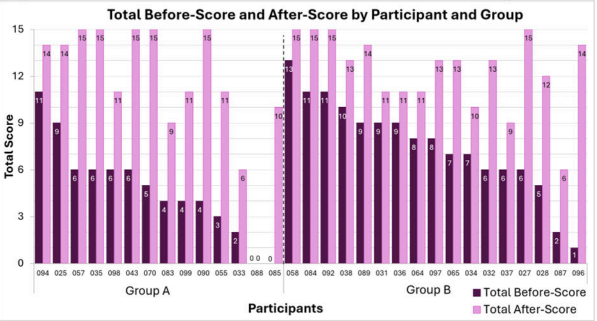


Figure 3: *Total before-score and after-score for each participant by group.* *<u>Before-score</u> shows a significant difference between groups (A mean = 4.7; B mean = 7.5; p=0.017). <u>After-score</u> shows no significant difference between groups (A mean = 11.5; B mean = 12.4; p=0.522).*

4.4. Between-sentence effects

Figures 4-6 show that a focus on average results masks considerable variation in responses to the original sentence uttered in each file:

1. It was a sunny day and the children were going to the park.
2. The camel was kept in a cage at the zoo.
3. The police returned to the museum.
4. The man read the newspaper at lunchtime.
5. He was sitting at his desk in his office.

Variation in before-scores is due in part to Group B's learning benefit creating higher scores for later files. Thus, while files 1 and 2 showed similar before-scores in both groups, scores for 4 and 5 were significantly higher for Group B. For Group A, by contrast, before-scores showed increasing scores of 0 for later files, suggesting these participants simply gave up trying.

However, both before- and after-scores suggest that files 1 and 5, perhaps also 4 to a lesser extent, are in some sense easier to understand than files 2 and 3. While files 1 and 5 were fully accepted by most (but not all) participants (87% and 84% respectively), only half fully accepted the suggested transcripts for Files 2 (52% overall) and 3 (48% overall).

At this stage it is not clear exactly what causes this differential effect. The first step for further exploration should be a follow-up study presenting the files in different orders.

4.5. Within-sentence effects

Further analysis revealed interesting differences in responses for individual key phrases within files. In this section, the number of correct responses is given as a percentage of all attempted responses (77% overall), setting aside those with 0 scores. The first observation (Figure 7) is that the end-phrase of each file was the most likely to be transcribed accurately (85% of attempted transcripts), while the accuracy for the start- (34%) and mid- phrases (37%) were notably lower. Thus 'the park' was accurately heard by 93% of those who attempted a transcript; 'the zoo' by 100%; 'the museum' by 86%; 'lunchtime' by 74%; and 'his office' by 63% (though this rises to 74% with the inclusion of 'the office' and 'office').

Again it is not entirely clear what causes this end-of-sentence advantage. It may simply be the end-focused intonation of these simple read sentences. Certainly more testing with a wider range of materials would be necessary before generalizing.

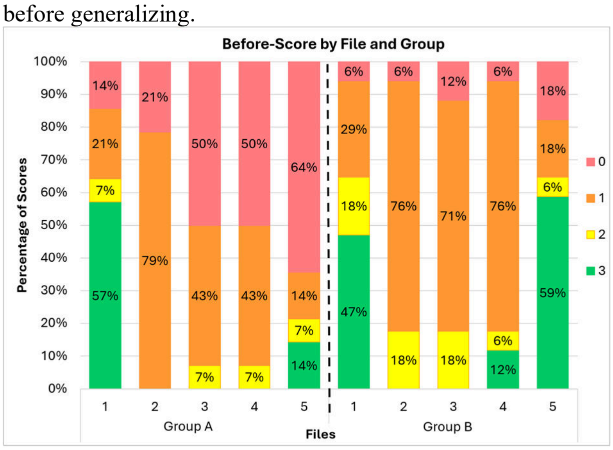


Figure 4: *Percentage of before-scores by file and group. Chi-square tests show that Groups A and B had similar scores for File 1 (p=0.669), File 2 (p=0.141), and File 3 (p=0.063); and significantly different scores for File 4 (p=0.031) and File 5 (p=0.04).*

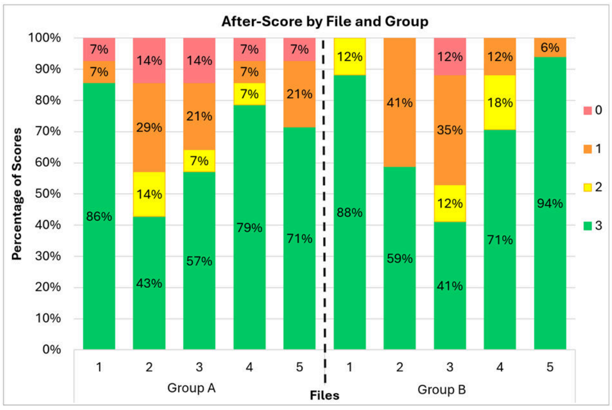


Figure 5: *Percentage of after-scores by File and Group. Chi-square tests by Group for each File show no significant difference (lowest p-value=0.1334).*

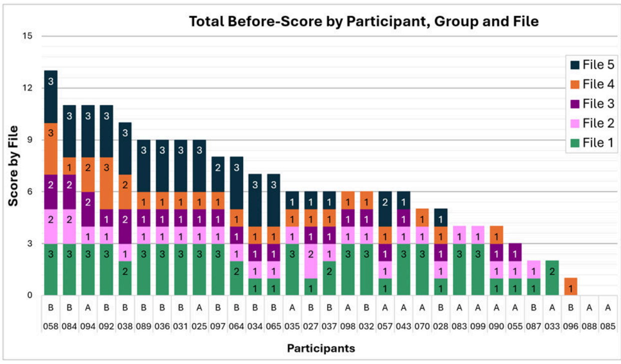


Figure 6: *Total before-scores for all 'L1' participants, shaded to show scores for each individual file.*

Particularly relevant in the present context is an analysis of common errors. Most notable is file 2, for which the correct transcription of the start-phrase ‘The camel’ was not given by any participants. Instead, three participants (11% of those attempting this file) gave ‘the owl’, ‘the owls’ or ‘owl/arrow’; three gave ‘now’, two gave ‘animal’, and one gave ‘cow’. In file 3, three (14%) gave ‘please’ as the start-phrase, while only two (9%) gave ‘police’. For file 4, the mid-phrase ‘the newspaper’ was transcribed correctly by six (26%), but three gave ‘museum’, two gave ‘music’ and one gave ‘muesli’.

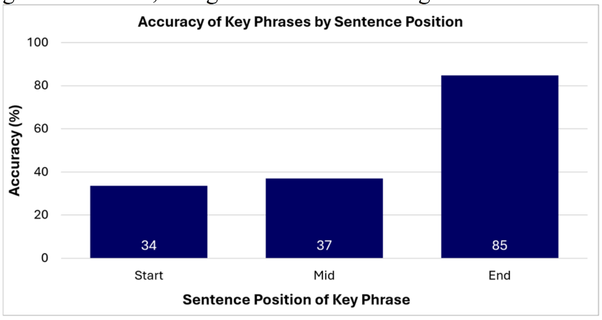


Figure 7: *Location of key phrases accurately transcribed before suggestion, as a percentage of attempted transcripts for each file.*

5. Discussion

Experiments with SWS and forensic transcription (FT) both yield observations of the strong power of a textual suggestion to make a meaningful utterance ‘pop out’ of unintelligible noise. However, while SWS research has focused on the *correct* utterance popping out, FT research has demonstrated that even a demonstrably inaccurate suggestion can create a confident pop-out perception. One of the difficult concepts speech scientists have to explain to the law is that just because listeners clearly hear particular words does not necessarily mean those words were the ones actually spoken [12]. As has long been known in phonetics, most stretches of speech are open to multiple interpretations even in clear speech (e.g. ‘grey day’ vs ‘grade A’) [13]; in poor-quality audio, interpretation depends heavily on listener expectations [14-15].

These considerations point to a major difference between SWS and FT research. All the results discussed above reflect scores predicated on ‘accuracy’ evaluated against a known ‘right answer’. This ‘ground truth’ of course is not available in FT, prompting the question of how we could evaluate the responses in the current experiment in its absence.

One common suggestion is to use artificial intelligence [see 16]. Running the five current files through Whisper gave results as shown (scores in brackets), for a total before-score of 8, lower than the top third (30%) of the L1 participants:

1. *It was a sunny day and the children were going to the park*. (3)
2. *The owl is kept in a cage in the zoo*. (2)
3. *Please visit us at the museum*. (1)
4. *Allow me to meet you for lunchtime*. (1)
5. *He was sitting with his girlfriend, so I laughed*. (1)

Another common suggestion is to evaluate responses against the acoustics of the file. Here, however, while acoustic analysis can rule out some responses on the basis of rhythm or other features, SWS does not allow clear evaluation at a segmental level (see Figure 8). This is also a problem with forensic audio, where the quality can make evaluation via acoustic analysis difficult [17]. Experts are often more reliable in ruling out an incorrect suggestion than in giving a demonstrably correct interpretation.

This highlights another important difference between SWS and FT: the nature of the audio files. SWS starts from clear, read sentences, produced by a single speaker, then degraded via a regular process. FT originates from unmonitored conversation, and is degraded via multiple, variable and unknown factors. Only in rare cases could we expect a rapid learning effect such as that seen with feedback in the above results (and see also [11] for a learning effect without feedback).

Of course the controlled situations of SWS and similar experiments have many advantages, and have given speech science much useful knowledge. Forensic transcription, however, offers challenges that are also worthy of exploration by speech scientists. These factors make it valuable for researchers to experience and study the process of transcribing forensic-like audio under a range of forensic-like conditions [18], and especially to explore the conditions which are liable to make listeners, including analysts, confident but wrong.

Finally, it is worth noting that the challenge of understanding indistinct speech of unknown content in potentially misleading contexts is exactly the challenge that listeners face in real-life situations every day. Researching FT therefore potentially has implications for theoretical accounts of human speech perception.

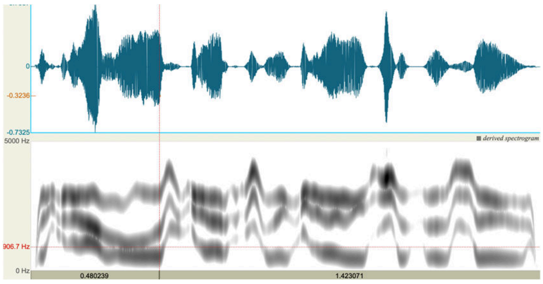


Figure 8: *Spectrogram of File 2, with location of ‘The camel’ indicated. While this is consistent with ‘the camel’, it could also support other interpretations.*

6. Conclusion

Experiments on sine-wave speech, like many other kinds of speech perception experiments, are typically carried out by researchers who know the true content of the audio, and are able to present systematically distorted samples to participants and systematically deprive them of relevant information. Such research can of course provide a great deal of useful knowledge about human speech perception. However it does have some risk of researchers getting ‘locked in’ to the ‘correct response’, without considering the validity of other responses, given the potential for the distorted acoustic signal to be interpreted in multiple ways.

The next step for this experiment is to continue investigating possible interpretations of sine-wave speech, with a particular focus on the effects of different feedback on the listeners’ interpretations (such as presenting the original sentence; a plausible but misleading sentence; or both of these options).

7. References

[1] Fraser, H. "The development of legal procedures for using a transcript to assist the jury in understanding indistinct covert recordings used as evidence in Australian criminal trials: A history in three key cases", Language and Law=Linguagem e Direito, 8(1), 59–75, 2021. doi: 0.21747/21833745/lanlaw/8_1a4

[2] Fraser, H. and Kinoshita, Y. "Injustice arising from the unnoticed power of priming: How lawyers and even judges can be misled by unreliable transcripts of indistinct forensic audio", Criminal Law Journal, 45(3), 142–152, 2021. http://hdl.handle.net/11343/285048

[3] Remez R., Rubin, P., Pisoni, D. and Carrell, T., "Speech perception without traditional speech cues", Science, 212(4497):947-950, 1981. http://www.jstor.org/stable/1685714

[4] Rubin, P. "SWS: an overview and history". Unpublished Manuscript, 2005. https://static1.squarespace.com/static/6463dc0cc0fa823cfe25e5fd/t/648bc8f2464bdb011edbfaeb/1686882549081/SWSOverview2005.pdf

[5] Barker, J. and Cooke, M. "Is the sine-wave speech cocktail party worth attending?", Speech Communication, 27:159-174, 1999, doi: 10.1016/S0167-6393(98)00081-8

[6] Davis, M. and Johnsrude, I. "Hearing speech sounds: Top-down influences on the interface between audition and speech perception", Hearing Research, 229(1–2):132–147, 2007, doi: 10.1016/j.heares.2007.01.014

[7] Lawton, G. "Mind tricks: Ways to explore your brain", New Scientist. 195(2622):34-41, 2007.

[8] Smith, C. and Critchlow, H. "Ketamine and schizophrenia", The Naked Scientists podcast, 2013. https://www.thenakedscientists.com/articles/interviews/ketamine-and-schizophrenia

[9] Seth, A. "Your brain hallucinates your conscious reality" TED talk. 2017. https://www.ted.com/talks/anil_seth_your_brain_hallucinates_your_conscious_reality/

[10] Fraser, H., Loakes, D., Knoch, U., and Harrington, L. "Towards accountable evidence-based methods for producing reliable transcripts of indistinct forensic audio", International Association of Forensic Phonetics and Acoustics (IAFPA). Zurich. 2023.

[11] Cooke, M., Scharenborg, O., and Meyer, B. T. "The time course of adaptation to distorted speech", The Journal of the Acoustical Society of America, 151(4), 2636–2646, 2022, doi: 10.1121/10.0010235

[12] Fraser, H. "'Assisting' listeners to hear words that aren't there: dangers in using police transcripts of indistinct covert recordings", Australian Journal of Forensic Sciences, 2018, doi: 10.1080/00450618.2017.1340522

[13] Lehiste, I. Readings in Acoustic Phonetics. MIT Press, 1967.

[14] Repp, B. "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception", Psychological Bulletin 92(1) 81-110, 1982, doi: 10.1037/0033-2909.92.1.81

[15] Shockey, L., and Bond, Z. "What slips of the ear reveal about speech perception". Linguistica Lettica, 22, 107–113, 2014.

[16] Loakes, D. "Automatic Speech Recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare?", Frontiers in Communication, 9(Capturing Talk), 2024, doi: 10.3389/fcomm.2024.1281407

[17] French, P., and Fraser, H. "Why 'ad hoc experts' should not provide transcripts of indistinct forensic audio, and a proposal for a better approach", Criminal Law Journal, 42(5), 298–302, 2018.

[18] Kettle, E. "Feel sorry for your ears": Exploring challenges in transcribing speech with unknown content in unfamiliar varieties. Australian Linguistics Society conference, 2024.

Applications of the Likelihood Ratio Framework in Forensic Speech Science Cases Involving Disputed Utterances, Tampering and Voice Lineups

Phil Rose

Emeritus Faculty, Australian National University
philjohn.rose@gmail.com philjohnrose.net

Abstract

Examples are given of the use of the likelihood ratio framework in real world case-work involving disputed utterances and tampering, and its theoretical application in voice lineups. Some of the problems in application are pointed out, especially with respect to estimation of priors.

Index Terms: likelihood ratio framework, Bayes' Theorem, disputed utterances, prior odds, tampering, voice lineups.

1. Introduction

About a quarter of a century has now passed since the first proposals to use likelihood ratios (LRs) in forensic speaker recognition. In that time it has become, after DNA, part of the new paradigm for the evaluation of forensic evidence [1] and now constitutes best practice for the *European Network of Forensic Science Institutes* [2: 7 *et pass*.]. However, forensic voice comparison, as it is now appropriately called, is not the only area of forensic speech science where evidence needs to be evaluated logically. The aim of this paper is to describe the use of LRs in three other areas of FSS: disputed utterance, tampering and voice-lineups – the first two from real case-work – and point out some of the problems involved.

The essence of the LR approach is to estimate the strength of evidence in favour of one hypothesis over another. For this one needs, of course, the two competing hypotheses and the evidence. For example, in forensic voice comparison the two hypotheses are often that the incriminating speech was said by the suspect; and that it was said by someone else. The evidence is usually some kind of quantification of the speech acoustics. The strength of the evidence is then straightforwardly the ratio of the probabilities of the evidence E under the competing hypotheses H_1, H_2: $p(E| H_1) / p(E | H_2)$. The LR is intended to indicate how much the evidence should make you rationally alter your belief in the hypotheses. Whatever your rational belief in a hypothesis before the evidence is adduced, the LR tells you how much you should strengthen or weaken that belief. The introduction of the notion of belief situates the LR in its proper place within Bayes' Theorem, where, in a mathematically simple and yet insightfully profound way, it provides the link between prior and posterior beliefs [3: 243, 4: 14-15].

2. Disputed Utterances

Did Abba really sing *see that girl, watch her scream, kicking the dancing queen*? Twenty-two percent of listeners apparently think so [5]. An inexhaustible list of hilarious mondegreens in many languages shows that people can mishear things when the sound is indistinct. Since recordings in real-world cases can get much less clear than pop-songs, it is sometimes disputed not *who* said something, but *what* was said. One of the best-known examples of disputed utterance comes from the New Zealand murder case *R v. Bain* where a prosecution witness claimed to hear Bain say the obviously incriminating *I shot the prick* in an emergency telephone call [6]. Alternative claims from defence experts were that Bain said *I can't breathe*, or *I can't help puking*, or *I can't touch it*, or that he was gasping for breath, or that it was not speech. This variety is what happens when you ask different unprimed people what they hear in a very indistinct recording. If you prime them, of course, you can, within reason, get them to hear what you want [7].

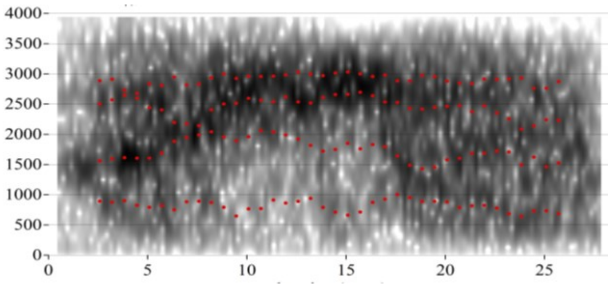


Figure 1: *Spectrogram with superimposed formants of* I shot/I can't *in R v. Bain. X-axis = duration (csec.), y-axis = spectral frequency (Hz)*

Perhaps the major benefit of a LR approach to disputed utterances is that it forces you to be precise about what is the evidence; and what the hypothesis. The temptation, of course, is to take what someone reports they heard as the evidence and then try to evaluate for example *what is the probability that x heard 'I shot the prick', given that that was indeed what was said*. But this is almost impossible to quantify. It is far better to take the observed acoustics as the evidence – that is, after all, the result of what was actually said – and the hypotheses as what someone reports they heard [8]. So the question becomes, for example: what is the probability of getting the observed acoustics, assuming that what was said was *I shot the prick*; and what is the probability of getting the observed acoustics, assuming that what was said was *I can't breathe*. Using the acoustics of the call as evidence, *not* what someone heard, has the major advantage that it is the same evidence for all hypotheses. The second major advantage is that the acoustics are quantifiable. In Bain it boiled down to a question of two hypotheses: the acoustics realize the /k/ in *can't* or the /ʃ/ in *shot*. Figure 1 shows the acoustics corresponding to *I shot* ~ *I can't*. The relevant portion can be seen to extend from about csec. 10 to csec. 15 and consists of narrowband noise between about 2 and 3 kHz for which *Praat* has extracted two resonances. It was pointed out in [8] that this portion neither sounds like, nor has the acoustic properties of the voiceless postaveolar fricative [ʃ] that would be expected from /ʃ/. Rather it sounds like, and has the acoustic properties of a voiceless palatal fricative [ç]. Bain can be heard elsewhere in the call to have a fronted velar [k̟] for his /k/, so one plausible explanation of the palatal fricative is that he was saying /k/, but failed to make the appropriate closure under the stress of coming across his all family dead. In other words, you would be more likely to get the observed acoustics if Bain had not said *shot*, but a word beginning with /k/. If Bain had not said 'I shot the prick' that does not of course imply his innocence, but he was eventually acquitted on other grounds. A subsequent experiment [9] demonstrated how to estimate the strength of evidence from the acoustics of [ʃ] and [ç], and that Bain's acoustics were vastly less likely if he had said [ʃ].

In order to actually estimate the posterior probability that Bain said *I can't* rather than *I shot* we would still need the priors with which to combine the LR. Given the enormous amount of available data that allow our mobile phones to predict rather well our next texted word, it would probably be possible to estimate the probability of *can't* occurring after *I* relative to the probability of *shot* in emergency phone calls.

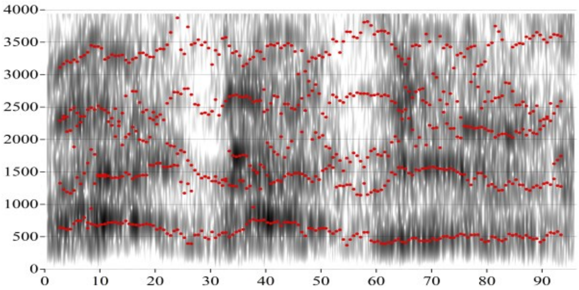


Figure 2: *Spectrogram with superimposed formants of chant 4*.

Normally a disputed utterance involves the speech of a single person. A case in October 2023, however, involved a whole crowd chanting during a pro-Palestinian protest in front of the Sydney Opera House. An accompanying video included captions stating *gas the Jews*! and went viral. The captioned video sparked understandable outrage and predictable schism and the media approached three forensic experts (I was one) [10]. It was mooted that the video chant might actually not be *gas the Jews!* but *where's the Jews?* In NSW, incitement to violence is an offence, and the police hired another forensic expert who concluded that the chant was 'with overwhelming certainty' *where's the Jews?* [11].

The fact that in this case we are talking about not one, but many people introduces several complications, e.g. that it would not be correct to persist with an all-encompassing hypothesis as *the crowd were chanting x*. The recording disseminated in the video would have picked up only a portion of the crowd, and – although that is not in the nature of a chant – a small number of them might have been chanting something different which by dint of their number was not picked-up (several individuals present reported hearing *gas the Jews*). There would have been other parts of the crowd that were not recorded simply because they were further away and out of range. It would therefore be more precise to formulate the competing hypotheses thus: *what was recorded on the disseminated video was* (1) *gas the Jews or* (2) *where's the Jews*.

In the captioned video played on her Sky News show by Sharri Markson as "proof" that *gas the Jews* was chanted [12], nine clear occurrences of the alleged chant can be made out. These are preceded by four full repeats of the chant *Allahu akbar* itself preceded by *takbiir!* (make great! – a conventional injunction to extoll Allah – by an individual leading the chant). As an example of Modern Standard Arabic the *Allahu akbar* chant is not phonetically accurate. For example the lateral seems to lack its appropriate gemination and pharyngealization; and the following long vowel should also have a pharyngealised (i.e low back) allophone. It is not possible to tell whether that is because some or all of the chanters were non-native speakers; or because of the act of chanting itself (c.f. the schwa vowel in *the* was lengthened to conform to the chant rhythm).

Figure 2 shows a reduced dynamic range spectrogram of chant number 4 with superimposed formants. Some of the formant structure has come out surprisingly clearly, which is remarkable, given that we are seeing the acoustics of many speakers. The segmental boundaries are not well-defined; neither is there any indication of putative /s/ or /z/ noise. It can be seen that both F1 and F2 in the first syllable have slightly lower values at their onset, but the F1 is subsequently at about the same height as the F1 in the *the*, whereas the F2 seems to have a slightly higher, rising trajectory than the F2 in *the*.

In order to get a handle on the probability of getting the questioned acoustics – in this case the F1 and F2 in the first word – under competing hypotheses of /we:/ in *where* and /æ/ in *gas*, we need to know what the F1-F2 monophthongal vowel space looks like. Thanks to the *Multicultural Australian English* project, we have F-pattern data for young speakers from several ethnically diverse sites in Sydney [13]. The site with the greatest connections to Islam is clearly Bankstown, with its ca. 17% Lebanese ancestry and ca. 21% of homes speaking Arabic as well as English [13]. The left panel of figure 3 shows an F1-F2 plot for the relevant vowels modified from [13]. It can be seen that the /e:/ has a very similar F1 location to the /ɜ:/, while having a higher F2; and that the /æ/ has a similar F1 extension to the /ɐ:/, while having an F2 a little higher than /ɜ:/.

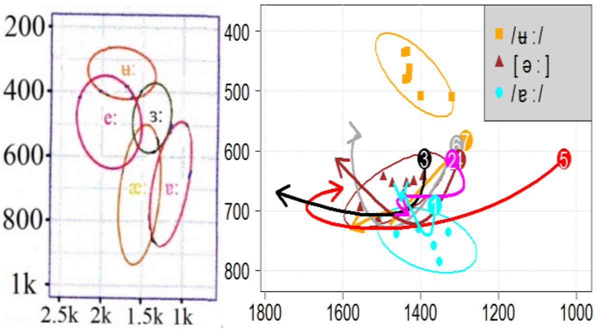


Figure 3: *Left = F1-F2 plot for young Bankstown male monophthongs (95% confidence ellipses). Right = F1-F2 trajectory of questioned syllables 1 to 7 plotted against known values. X-axis = F2, y-axis = F1.*

The right panel of figure 3 shows plots of the F1-F2 trajectory of questioned syllables in chants 1 to 7 against an F1-F2 plot of all the long chanted vowels (/u:/ in *Jews*, [ə:] in *the*, /ɐ:/ in *Allahu akbar*). (The questioned vowel in the last two chants was obscured.) It can be seen that the trajectories involve glides from a retracted (low F2) position towards a space that would be expected for /e:/. Although their quantification is not clear, these characteristics would be very much more likely if the questioned word had been *where* rather than *gas*, which would be expected to have a short backing and falling perturbation from the velar stop. Assuming flat priors would mean that the quantifiable chants were very likely not *gas the Jews*. However, unlike perhaps in Bain, it is not at all clear how one would estimate the priors in this case in order to arrive at the posterior probability quoted by the police expert.

3. Tampering

Sometimes it is claimed that a recording has been tampered with in some way. Perhaps something incriminating has been removed or inserted. In a 2000 case in the UK [14] it was alleged that the phrase *the… Law Society* (… indicates a pause) had been copied from the utterance *the … Law Society ringing me?* in an earlier part of a recording and pasted into an incriminating location later in the recording. An audio engineer retained by the prosecution noted the high degree of visual similarity in waveform between the two occurrences of *the… Law Society* and, adducing the phonetic maxim that one never says the same thing in an identical way, opined that the second occurrence had indeed been copied. Looking at the spectrograms and F0 of the two utterances in figure 4, one can have some sympathy with the engineer. Any phonetician, too, with experience in the variability of natural speech would think that kind of near-identity unlikely if the utterance had been repeated naturally. Phoneticians retained by defence argued, however, that the phonetic maxim upon which the engineer relied "had not been proved" and that the second occurrence of *the Law Society* had just been said in a very similar way to the first. One of the judges pertinently drew attention to the lack of quantification, asking: "… to what extent can the differences between the same word, spoken by the same person, be measured? Does it follow that where the word is electronically measured and compared to another word and there are marked similarities that must mean it is a digital clone?"

Indeed! So, how would one evaluate these claims logically, and quantitatively, within a LR framework? Call the first occurrence LS1 and the alleged copy LS2. The alternative hypotheses are (H_c) that LS2 has been copied and pasted from LS1; and (H_n) that LS2 does not so originate. The evidence E_a is the difference between the acoustics of LS1 and LS2. In a LR framework what has to be evaluated is the probability of getting E_a assuming that LS2 has been copied from LS1, and assuming that it has not. Note that this is not the question asked by either judge or engineer: they were typically asking about the probability of a hypothesis (digital cloning), given the high degree of similarity (evidence).

Since there is an infinite number of parameters one could use to estimate E_a, it is sensible to simplify. F0 was chosen as a suitable parameter to compare: it is easy to measure and its perceptual correlate of pitch is more straightforward to imitate than segmental and voice quality. When aligned with respect to the details of the waveform at the onset of the F0 in *Law*, the mean absolute difference between the F0 of the two utterances was 0.9 Hz. This is taken as E_a.

An experiment was then run to see how likely this value of 0.9 Hz is under competing hypotheses: when the phrase *the … Law Society* is digitally copied; and when it is repeated. The top panel of figure 5 shows the F0 of 30 attempts by me, not to just repeat, but actually imitate both intonational pitch and pause of *the…Law Society* as said in the recording. It can be seen that the F0 over the *the* shows a small amount of variation, but obviously getting the duration of the pause right introduces considerably more. Each of my 30 imitations was then digitally copied in *Praat* to different random downstream locations and the difference measured between original and copied F0.

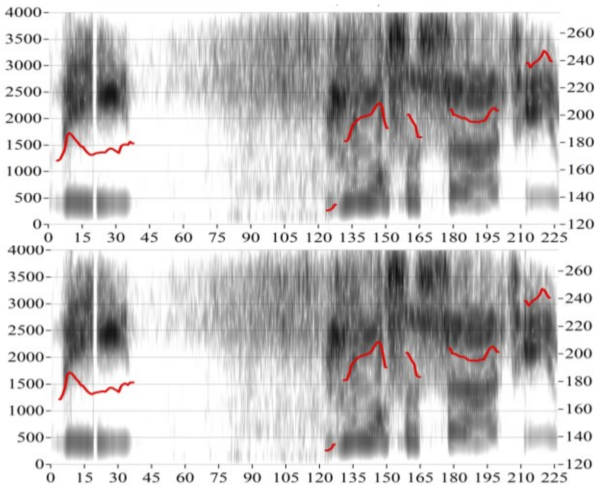


Figure 4: *Acoustics of* the…Law Society. *Top = original, bottom = alleged copy. X-axis = duration (csec.), y-axes: left = spectral frequency (Hz), right = F0 (Hz).*

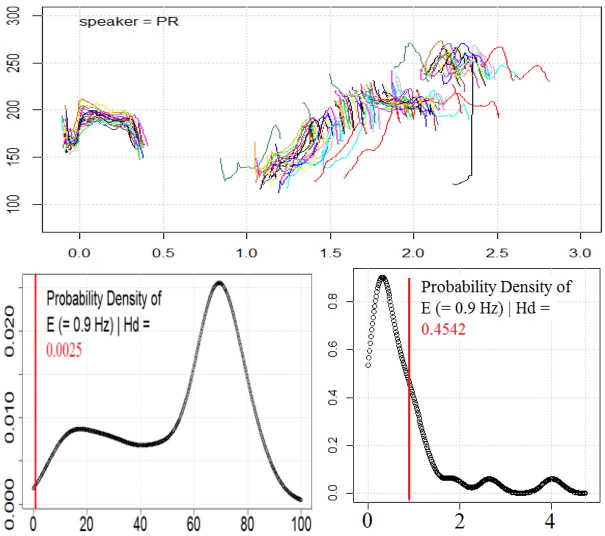


Figure 5: *Determining LR in tampering case. Top = F0 of 30 attempts to imitate* the…Law Society. *Bottom = distributions of absolute F0 differences between different (left) and copied imitations. Vertical red line = evidence.*

The bottom left panel of Figure 5 shows the probability density for all 435 pairwise differences in F0 between the imitated repeats. It shows that it is indeed possible to get an almost exact repeat – where the difference is close to zero – but most of the time the difference is much bigger (the reason for the distributional bimodality is not clear). As can be seen, the probability density of getting the value of 0.9 Hz for the evidence is a relatively low 0.0025. The bottom right panel of figure 5 shows the probability density for the digitally copied pairs. It is not, as one might at first expect, zero – there is for example one occurrence of a difference as big as 4 Hz. But of course the probability of getting the evidence assuming copying is relatively quite high: 0.4542. The difference between LS1 and LS2 is therefore (0.4542/0.0025 =) 180 times more likely if LS2 had been copied. This magnitude of LR already has some heft. But it will be much greater because the defence hypothesis was that it was a case of natural occurrence, not imitation. But there is still more, by way of contribution from the spectral properties. At the end of the spectrogram in the top panel of figure 4 can be seen the well-known anticipatory drop in F2 and F3 in the F-pattern for /i:/ in *society* caused by the following post-alveolar approximate [ɹ] in *ringing*. But the same perturbation can also be seen in the alleged copy, where *society* is not followed by an /r/ and there is therefore nothing to cause such a perturbation. The LR for this must be very big. The probability of getting the perturbation in /i:/ F-pattern given that it is followed by [ɹ] must approach 1. The probability of getting the perturbation with nothing following to condition it must be very low. (A devil's advocate might imagine a scenario where a speaker intends to say a word beginning with /r/, but then for some reason stops.) Perhaps one could estimate this probability by scouring recordings of natural speech for disfluencies of this type and estimate the amount of occurrences per unit time. But there is yet more: in figure 4 a dropout is evident in *the* in both LS1 and LS2 with the same duration and at precisely the same place. Again the LR for this must be vanishingly small. We now encounter a problem. Clearly, combining the LRs for the F0, the F-pattern and the dropout will result in an LR of enormous magnitude, but we cannot calculate it. Neither is it clear how to estimate prior odds for this scenario.

4. Voice Lineups

During an armed robbery someone hears the disguised robber speak. A suspect is arrested and the earwitness identifies the suspect in a voice lineup. Assuming the lineup was conducted fairly, what evidential strength should one give that identification? Is that good news for the prosecution?

A lot of empirical work has been done recently by the members of the UK project *Improving Voice Identification Procedures* (IVIP) to determine optimum parameters and procedures for voice lineups – for example, whether to tell the earwitness that the suspect might not actually be in the lineup; what the optimum number of foils should be; or how long the voice sample [15]. But surprisingly almost nothing has been done on arguably the more important question: how to evaluate the strength of evidence of a voice lineup identification, that is, where the earwitness picks out the suspect. Interestingly, attempts to address the question in visual line-ups e.g. [16] appear to have focused on signal detection parameters familiar in automatic speaker recognition like DET and ROC, but which relate to posterior probabilities, not the strength of evidence. In 2002, in an attempt to explain Bayes' Theorem to the legal profession, a judge of the Supreme New South Wales Court of Appeal who has also published widely on probability showed that the LR in a voice lineup is a function of the reliability of the earwitness and the number of foils; and how identification through a properly conducted lineup carries more evidential strength as the number of foils increases [17]. The argument below, presented in [18], is from that paper. A much later paper [19] has also approached the problem using Bayes Factors.

We assume the suspect is present in a lineup and the earwitness picks them out. In this scenario, the evidence E_s is the fact that the earwitness picked the suspect. What is the value of that evidence? The prosecution hypothesis (H_p) is that the suspect is the person the witness heard; the defence hypothesis (H_d) is that the suspect is not the person the witness heard. The strength of evidence will be the ratio of the probabilities of the evidence E_s under the competing hypotheses, viz: LR = $p(E_s \mid H_p) / p(E_s \mid H_d)$.

In order to estimate the probability that the witness would pick the suspect assuming the suspect was whom the witness heard we really need to know how good the earwitness actually is. Of course the best option is to test the witness; but *faute de mieux* one can estimate this from known naïve identification under circumstances similar to the crime. In [17] the hypothetical witness was assumed, completely unrealistically, to be 90% reliable. Experiments with simulated voice lineups in the *Improving Voice Identification Procedures* (IVIP) project report rates between ca. 45% and 30% when the suspect was present in the lineup. IVIP felt that, although this is poor, in reality the performance would be better, since real lineups would be more heterogeneous. Earlier experiments with naïve unfamiliar listeners have indeed shown better figures, but not that much better: between 40% and 50% [20: 202, 203]. Since the IVIP experiments were conducted under differing conditions to find optimum parameters, it makes sense to use their best performance – about 45% with 15 second duration speech and nine voices in the lineup – because the worst performances could have been caused by experimental choice of an adverse parameter. Assuming the witness is 45% reliable will give the value 0.45 for the probability of the witness picking the suspect assuming they had in fact heard them.

Estimating the other part of the LR – the probability that the witness would wrongly identify the suspect as the person they heard – can be broken down into two questions. We have to ask firstly, and obviously, what is the probability that the selection is wrong (i.e. the suspect is not whom the witness heard). With the earwitness' 45% reliability, this is 0.55. We also have to ask what is the probability that the suspect is selected. IVIP concluded that the nine-voice parade mandated in the UK should be maintained, so this is (1/9 =) 0.11… . The probability that the suspect is selected, and they are not whom the witness heard, is then (0.55 * 0.11… =) 0.061… . Thus the LR for this identification would be (0.45 / 0.061… =) ca. 7.4. Given the generally poor performance of naïve unfamiliar recognition this might surprise some, but as emphasised in [17] it is also a function of the number in the lineup, and shows the benefit of an explicit, quantifiable LR approach. But 7.4 is still not very strong evidence. In the absence of other evidence pointing to the suspect as the perpetrator, the biggest posterior probability achievable with a LR of 7.4 (with even priors, i.e. with only two possible people who could be guilty) is about 88%. But of course in conjunction with other available evidence it can be a powerful multiplier.

5. Summary

This paper has given an idea of what is involved in applying the likelihood ratio framework to cases of disputed utterances, tampering, and to voice lineups. In the disputed utterance cases it was shown that the evidence is the acoustics; not what someone reports hearing. In the tampering case there were potential problems with knowing the LR was inconceivably big but being unable to calculate it. Priors, already characterized in [21: 88] as an 'open-ended problem', were also seen to be a potential problem in both disputed utterance and tampering. Perhaps one can avoid this in tampering and disputed utterances with uninformative priors; perhaps not.

6. Acknowledgements

This paper builds on my keynote at the 2021 conference of the International Association of Forensic Phonetics & Acoustics. I thank Prof. Cox for providing from a preprint the indispensable data on Bankstown formants, and Prof. McDougall for providing references on evaluation of visual lineups. Many thanks also to my anonymous reviewers for their very useful comments, most of which I was able to address.

7. References

[1] Morrison, G. S., “Forensic voice comparison and the paradigm shift”, Science & Justice 49: 298-308, 2009.

[2] Drygaylo, A., Jessen, M., Gfroerer, S., Wagner, I.,Vermeulen, J. and Niemi, T., “Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition”, Verlag für Polizeiwissenschaft, 2015.

[3] Silver, N., The Signal and the Noise - The Art and Science of Prediction, Penguin, 2013.

[4] Ellenberg, J., How Not to be Wrong, Penguin, 2015.

[5] https://www.newsweek.com/50-famously-misheard-misunderstood-song-lyrics-explained-1561728

[6] Innes, B., “R v David Bain – A unique case in New Zealand legal and linguistic history”, International Journal of Speech, Language and the Law, 18: 145–155, 2011.

[7] Fraser, H., “Enhancing’ forensic audio: what if all that really gets enhanced is the credibility of a misleading transcript?”, Australian Journal of Forensic Sciences, 52(4): 465-476, 2019.

[8] Rose, P., “Evaluation of Disputed Utterance Evidence in the matter of David Bain’s Retrial”, http://philjohnrose.net/pubs/FVC_pubs/index.html, 2009.

[9] Morrison, G. S. and Hoy, M., “What did Bain really say? A Preliminary Forensic Analysis of the Disputed Utterance Based on Data, Acoustic Analysis, Statistical Models, Calculation of Likelihood Ratios, and Testing of Validity”, 46th AES International Conference,

[10] Wilson, C., and Lattouff, A., “New footage and audio experts raise further doubts about Sydney Opera House protest video”, https://www.crikey.com.au/2023/12/19/new-footage-audio-experts-sydney-opera-house-protest-video/ 2023.

[11] https://www.abc.net.au/news/2024-02-02/nsw-police-opera-house-protest-video-analysis/103418582

[12] https://www.youtube.com/watch?v=XGdM6FcSYXw

[13] Cox, F. and Penney, J., “Multicultural Australian English The new voice of Sydney”, Australian Journal of Linguistics 1-20, https://doi.org/10.1080/07268602.2024.2380680, 2024.

[14] Rose, P**.,** **“**Forensic speech science report in a case involving the alleged tampering of a recording**”,** http://philjohnrose.net/pubs/FVC_pubs/index.html , 2009.

[15] McDougall, K., Nolan, F., Paver, A., Smith, H., Braber, N., Wright, D., Pautz, N., Goodwin., P., Robson,. J. and Mueller-Johnson, K., “Improving Voice Identification Procedures: Findings and Recommendations from the IVIP Project”, IAPFA Webinar 18th April 2024.

[16] Wixted, J.T., and Mickes, L., “Evaluating eyewitness identification procedures: ROC analysis and its misconceptions”, J. Applied Research in Memory and Cognition, 4(4): 318-323, 2015.

[17] Hodgson, D., “A Lawyer looks at Bayes’ Theorem”, The Australian Law Journal, 76: 109-118, 2002.

[18] Rose, P., “Evaluating Strength of Evidence in Voice Lineups”, Symposium to honour Andy Butcher, Melbourne University, http://philjohnrose.net/pubs/FVC_pubs/index.html, 2017.

[19] Rosas, C., Sommerhoff, J. and Morrison, G.S., “A method for calculating the strength of evidence associated with an earwitness’s claimed recognition of a familiar speaker”, Science & Justice, 59: 585-596, 2019.

[20] Hollien, H., The Acoustics of Crime, Plenum, 2002.

[21] Jaynes, E.T., Probability Theory – The Logic of Science, CUP, 2003.

Locating Spectral Regions of Speaker Sensitivity with the Sub-Band Mel-Frequency Cepstrum: An Investigation Guided by Forensic Voice Comparison

Frantz Clermont, Shunichi Ishihara

Speech and Language Laboratory, Australian National University, Canberra, Australia
dr.fclermont@gmail.com, shunichi.ishihara@anu.edu.au

Abstract

The non-uniform spectral encoding of speaker information suggests that certain sub-band regions carry strong sensitivity to that information. Such regions are sought here by: capturing locally-encoded spectral information with band-limited cepstral coefficients (BLCCs), and quantifying the sensitivity through forensic voice comparison (FVC). Using 5 Japanese vowels from 306 native male speakers, FVC experiments were conducted with BLCCs representing 18 sub-bands spaced across the full range [0-4000 Hz] on the mel scale. The FVC results highlight 3 speaker-sensitive sub-bands, located near formant regions in low-, mid-, and high-frequency ranges. Speaker information is strongly encoded in high-frequency sub-bands roughly above 2300 Hz.

Index Terms: Band-limited cepstral coefficients (BLCCs), mel-frequency scale, filter-bank, vowel, speaker sensitivity.

1. Introduction

Ever since the influential paper [1] which reported the superior performance of mel-frequency cepstral coefficients (MFCCs) in speech recognition, it has been argued that these acoustic parameters based on a logarithmic scale possess the significant advantage of increasing resolution in the lower frequency bands, while allowing "better suppression of insignificant spectral variation in the higher frequency bands" [1: 364]. MFCC spectra are thus enhanced in the lower bands that encode the bulk of phonetic differences. Yet, MFCC spectra have also been shown to perform well in speaker recognition [2,3].

The following question naturally arises: Which regions of MFCC spectra encode the bulk of speaker differences? We investigate this question with: (1) sub-band MFCCs to focus on local regions, one at a time, across the full band; and (2) forensic voice comparison (FVC) to quantify speaker sensitivity in each selected sub-band. This approach is put forward as a key to finding speaker-sensitive sub-bands, and potentially assisting forensic speech experts for a deeper analysis of FVC outcomes and a more logical validation of the adopted FVC system.

Sec. 2 describes the method [4] recently developed for obtaining sub-band MFCCs, also referred to as band-limited CCs (BLCCs) in this work. By contrast with the alternative method of repeating the spectrum-to-cepstrum conversion for every sub-band, we have adopted the BLCC method which can easily transform full-band into sub-band MFCCs with flexible sub-band selection. The sub-band results presented in Sec. 4 may thus be seen as supportive evidence for the BLCC method.

Sec. 3 outlines our FVC experiments and multi-speaker vowel data. Using 18 sub-bands spaced across the full band [0-4000 Hz] on the mel scale, the experiments were aimed at locating the vowel spectral regions most sensitive to speaker differences with FVC performance as a quantitative measure.

Sec. 4 gives a detailed map of the least and most speaker-sensitive sub-bands. The top 3 optimal sub-bands and the full band are also compared in terms of FVC performance. Sec. 5 discusses the main findings with suggestions for further study.

2. The BLCC method

Sec. 2.1 gives an overview of the method. Sec. 2.2 outlines the mathematical formulae, and Sec. 2.3 discusses the practical size for a BLCC vector.

2.1. Procedural steps

The BLCC method proceeds in 3 main steps shown in Fig. 1. Steps (1) and (2) describe standard procedures of spectral analysis applied to short-time frames of the speech signal with a sampling frequency F_s (Hz). The final step (3) consists of a linear transformation from full-band to sub-band cepstrum.

At Step (1), the log magnitude spectrum (LMS), $S(\omega)$, is computed from the energy outputs of a filter bank. The filters' centre frequencies are non-uniformly spaced across the full range $[0, (F_s/2)]$ in Hz (or $[0, \pi]$ in radians). The spacing follows the mel scale, roughly linear at low frequencies and nonlinear at high frequencies.

At Step (2), the Discrete Cosine Transform (DCT) is used to decorrelate $S(\omega)$ and reduce its dimensionality. The result is a Fourier series of cosine functions weighted by the so-called cepstral coefficients C_k. In this work, the C_k are referred to as full-band, mel-frequency CCs (MFCCs in short). The average of the full-band LMS is usually assumed to be zero, hence $C_0 = 0$. In practice, the series is truncated after $M \ll \infty$ terms:

$$S(\omega) \cong \sum_{k=1}^{M} C_k \cos(k\omega), \quad 0 \le \omega \le \pi \qquad (1)$$

Independently of the spacing on the frequency axis, the retention of a small number of terms carries a smoothing effect of its own, one which makes $S(\omega)$ more immune to "noninformation bearing variabilities" [5: 169].

At Step (3), BLCCs are obtained using a new method which gives the flexibility of selecting a sub-band region of the full-band spectrum *without having to repeat the previous steps*. As shown in [4], the vector $\mathbf{c}'$ of BLCCs representing a sub-band can indeed be calculated *directly* from the vector $\mathbf{c}$ of full-band C_k using a linear transformation $\mathbf{A}$ expressed in Sec. 2.2.

2.2. Linear transformation formulae

A sub-band region $[\omega_1, \omega_2]$ of the cepstrally-smoothed, full-band LMS is represented by the Fourier cosine series given in Eq. (2). The mathematical aim is to express the band-limited coefficients C'_l as functions of the full-band C_k.

$$S(\omega(\omega')) \cong C'_0 + \sum_{l=1}^{N} C'_l \cos(l\omega'), \ 0 \le \omega' \le \pi, \qquad (2)$$

where C'_l is the l-th BLCC and N is the series' upper bound.

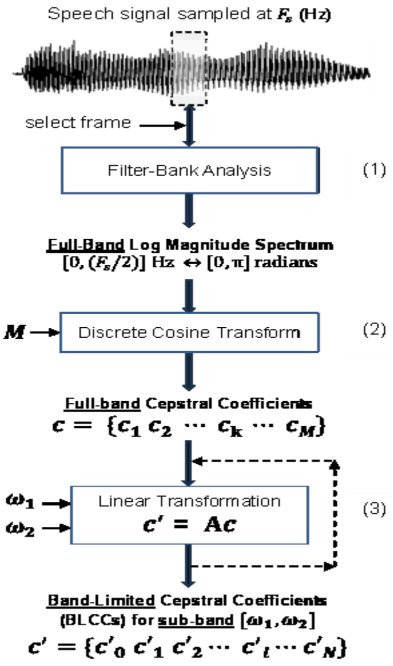

Speech signal sampled at F_s (Hz)
select frame
Filter-Bank Analysis
(1)
Full-Band Log Magnitude Spectrum
$[0, (F_s/2)]$ Hz ↔ $[0, \pi]$ radians
M
Discrete Cosine Transform
(2)
Full-band Cepstral Coefficients
$c = \{c_1\ c_2\ \cdots\ c_k\ \cdots\ c_M\}$
ω_1
ω_2
Linear Transformation
$c' = Ac$
(3)
Band-Limited Cepstral Coefficients
(BLCCs) for sub-band $[\omega_1, \omega_2]$
$c' = \{c'_0\ c'_1\ c'_2 \cdots\ c'_l \cdots\ c'_N\}$

Figure 1: *The BLCC method and its main steps.*

Note that Eq. (2) includes C'_0 because the average of $S(\omega(\omega'))$ within a sub-band may not be zero. The other $C'_{l>0}$ account for the spectral shape within the selected sub-band.

The frequency variable ω' defined below translates the sub-band interval $[\omega_1, \omega_2]$ to that of the full-band range $[0, \pi]$:

$$\omega' = \pi\left[\frac{(\omega-\omega_1)}{(\omega_2-\omega_1)}\right],\ \omega_1 \le \omega \le \omega_2 \qquad (3)$$

From Eq. (3) the frequency variable ω of the full-band series can be expressed as follows, where the scalar W is the ratio of the sub-band's width to the full-band's frequency range:

$$\omega(\omega') = \omega_1 + \left[\frac{(\omega_2-\omega_1)}{\pi}\right]\omega' = \omega_1 + W\omega' \qquad (4)$$

The notation $\omega(\omega')$ is a reminder that ω is a (band-dependent) function of ω'. It is thus possible to substitute ω in Eq. (1) for Eq. (4), and use standard formulae to obtain a *linear* relation between the band-limited C'_l (Eq. 2) and full-band C_k:

$$C'_l = \sum_{k=1}^{M} a_{lk}\, C_k,\ l = 0,1,\dots,N \qquad (5)$$

Eq. (5) leads to the matrix form $\mathbf{c}' = \mathbf{A}\mathbf{c}$, where $\mathbf{A}$ is the transformation matrix for the selected sub-band, $\mathbf{c}'$ the column vector of C'_l, and $\mathbf{c}$ the column vector of C_k. The elements of $\mathbf{A}$ are defined in Eq. (6a) for $l = 0$ and Eqs. (6b)-(6c) for $l > 0$:

$$a_{lk,\, l=0} = \beta_k[\sin(k\omega_2) - \sin(k\omega_1)] \qquad (6a)$$

$$a_{lk,\, l\neq kW} = \gamma_{lk}[(-1)^{l+1}\sin(k\omega_2) + \sin(k\omega_1)] \qquad (6b)$$

$$a_{lk,\, l=kW} = \cos(k\omega_1) \qquad (6c)$$

where:

$$\beta_k = \frac{1}{k(\omega_2-\omega_1)} \text{ and } \gamma_{lk} = \frac{2(kW)}{\pi[l^2-(kW)^2]} \qquad (6d)$$

2.3. BLCC vector: Practical size and spectral resolution

The Fourier-series (BLCC) representation of a sub-band region will theoretically improve with increasing values of the upper bound N. How large does N need to be in practice? The solution proposed in [4] is to truncate the BLCC series after $N = M \times W$ (MW in short) terms: M is the size of the full-band vector of C_k, and W (defined earlier) is the fraction of the full-band's range occupied by the sub-band's width. Note that MW will generally not be an integer; it must therefore be rounded for practical use.

Our rationale for MW is this: If M cepstral coefficients represent the full-band spectrum with a certain resolution, then roughly the same resolution for a sub-band region should be achievable with $C'_{l=0,1,\ldots N=MW}$. That is, for N fixed at around MW, there should be no significant loss in spectral resolution. The numerical profiles of BLCCs illustrated in [4,6] show that the dominant coefficients indeed extend up to MW, followed by a noticeable decay of their magnitude towards zero.

3. Experimental procedure

3.1. Speech materials and parametrisation

The speech materials were sourced from the Japan's National Research Institute of Police Science (NRIPS) database [7], consisting of microphone recordings from 306 adult-male, native Japanese speakers (mean age: 39.9, SD: 15.5). The 5 Japanese vowels (/a, e, i, o, u/ in null consonantal context) selected for this work were uttered twice in 2 sessions split 3-5 months apart. A sampling rate of 8000 Hz was used to contain the full frequency range within the Japanese telephone bandwidth [0-4000 Hz]. This bandwidth constraint combined with the non-contemporaneous recordings satisfy some of the basic requirements for forensic relevance.

Using 25 triangular-shaped filters spaced evenly on the mel scale with 50% overlap, filter-bank analysis was applied to each vowel segment using a frame size of 25 msecs and a step size of 5 msecs. For each frame, 14 MFCCs were obtained by performing DCT on the filter's log energy outputs. The MFCCs averaged over the 3 middle frames were retained as the vector of full-band MFCCs.

Subsequently, the full frequency range [0-4000 Hz] was scanned with a 600-Hz sub-band, shifted every 200 Hz, and the vector of BLCCs was obtained for each sub-band by linear transformation of the full-band MFCCs as described in Sec. 2. There are 18 sub-bands in total, resulting in 18 corresponding vectors of BLCCs. Applying the formula in Sec. 2.3, MW equals 2.1, rounded to 2. The size of each BLCC vector is therefore 3, including the 0th-order coefficient.

3.2. Data partitioning and experimental descriptions

The speech materials were randomly divided into 3 batches, which were rotated and used as the test, background, and calibration databases, resulting in 6-fold cross-validation FVC experiments (see Table 1). The results presented in this paper are the averages of these 6 experiments. There are 612 same-speaker comparisons and 61,812 different-speaker comparisons from these cross-validation experiments.

Table 1: *Six-fold cross-validation experiments*

	Test	Background	Calibration
1	Batch 1	Batch 2	Batch 3
2	Batch 1	Batch 3	Batch 2
3	Batch 2	Batch 3	Batch 1
4	Batch 2	Batch 1	Batch 3
5	Batch 3	Batch 2	Batch 1
6	Batch 3	Batch 1	Batch 2

Table of Contents for this manuscript

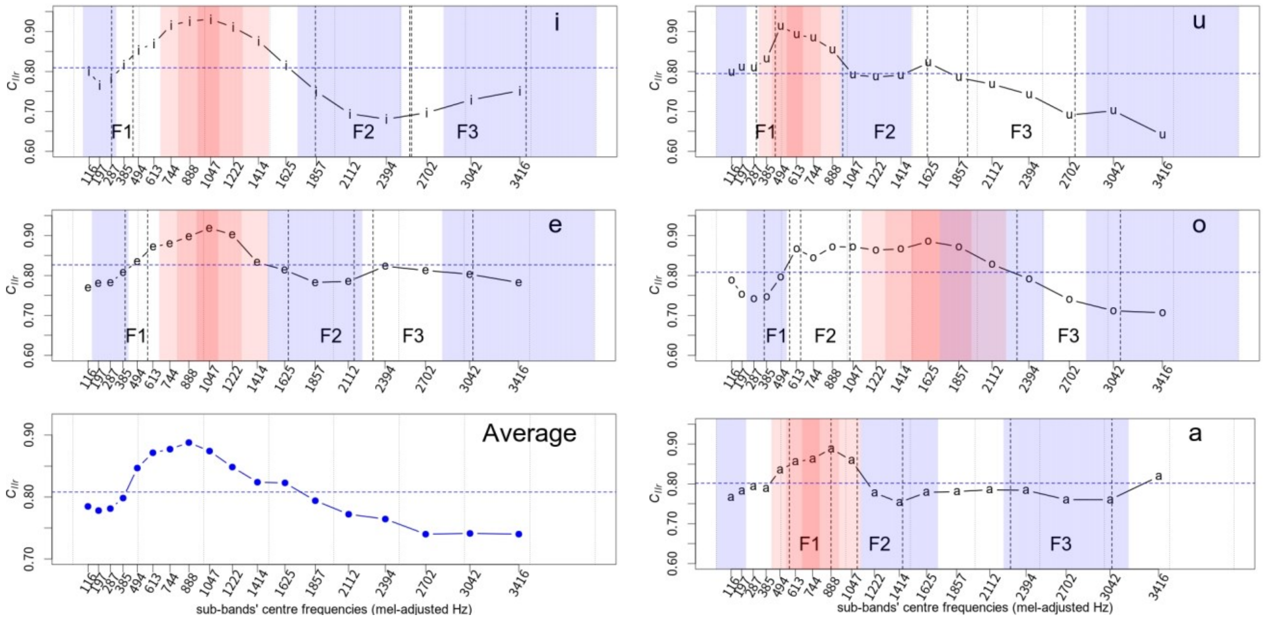


Figure 2: *Results for each of the 5 Japanese vowels and for all vowels are displayed in separate panels: C_{llr} values (y-axis) for the 18 sub-bands (x-axis) are connected by a black curve, and the vowel-averaged C_{llr} values by a blue curve. In every panel, the horizontal dashed line (in blue) indicates the C_{llr} value averaged over the 18 sub-bands. The 3 best-performing sub-bands (fused) are filled in blue, while the 3 under-performing sub-bands (fused) are filled in pink. In each vowel panel, the vertical dotted lines mark F1, F2 and F3 regions ±1 SD away from the speaker-averaged means.*

3.3. Experimental descriptions

Two FVC experiments were conducted on a per-vowel basis, with BLCCs representing each of the 18 sub-bands. Likelihood Ratios (LRs) were obtained using the Multivariate Kernel Density model [8], followed by logistic regression calibration [9]. The performance of the FVC system was assessed using the log-LR cost (C_{llr}) for each sub-band. *The closer the C_{llr} is to 0 and below 1, the better the performance of the FVC system*.

Experiment 1 sought to uncover how speaker information is encoded across different spectral regions for each vowel. If a sub-band is more sensitive to speaker information, the FVC system performs better with its BLCCs than with those from other sub-bands. Experiment 2 systematically combined sub-band LRs from Experiment 1, and compared the performance with full-band MFCCs results to locate the optimal sub-bands.

3.4. Sub-band limits on mel-frequency scale

Although the 18 sub-bands were selected on an equally-spaced Hz scale to maintain the uniform analytical potential of each sub-band, their corresponding C_{llr} values need to be positioned with respect to the non-linear scale on which the full-band MFCCs are based. To this end, the frequency axis of the C_{llr} graphs was mapped onto a mel-adjusted Hz scale. The adjusted limits $[\omega_1, \omega_2]$ for the 18 sub-bands are given in Table 2.

Table 2: *Sub-bands' mel-adjusted limits (Hz).*

1	[0, 231]	2	[70, 324]	3	[147, 427]
4	[231, 539]	5	[324, 663]	6	[427, 799]
7	[539, 949]	8	[663, 1114]	9	[799, 1295]
10	[949, 1494]	11	[1114, 1714]	12	[1295, 1955]
13	[1494, 2220]	14	[1714, 2511]	15	[1955, 2832]
16	[2220, 3185]	17	[2511, 3573]	18	[2832, 4000]

This adjustment is also desirable because it facilitates a more intuitive interpretation of sub-band locations, and a direct comparison with analogous results from previous studies.

4. Results and discussion

The results of Experiments 1 and 2 are depicted jointly in Fig. 2, but discussed separately in Secs. 4.1 and 4.2, respectively. The F1, F2, and F3 regions (means ±1 SD) for each vowel are marked by vertical dotted lines. Formant information is provided for reference only. See [10] for details on the formant measurement procedure.

The C_{llr} values (y-axis) from Experiment 1 are plotted in Fig. 2 against the sub-bands' centre frequencies (x-axis) for each vowel (see panels labelled "i", "u", "e", "o", and "a"). The C_{llr} values averaged over all vowels are shown in the panel labelled "Average". The average C_{llr} value of the 18 data points in each panel is shown as a dashed horizontal line to help the visual comparison of C_{llr} fluctuations from vowel to vowel.

Note that the centre frequencies are the midpoints of the sub-band limits given in Table 2. Following the non-linear property of the mel scale, the spacing between the 18 data points plotted in Fig. 2 becomes wider as the frequency increases towards the higher spectral regions.

4.1. Experiment 1

All C_{llr} values in Fig. 2 remain consistently below 1, suggesting that every sub-band carries some useful speaker information. The fluctuations in C_{llr} values clearly indicate that speaker information is non-uniformly distributed across the full band, a phenomenon observed in previous studies [11-15]. Owing to its flexible sub-band selection and computational efficiency, our BLCC method has facilitated the task of creating the detailed map of speaker-sensitive sub-bands superposed in Fig. 2.

The C_{llr} patterns in the vowel panels of Fig. 2 are consistent as far as relating local extrema to sub-band regions that are more or less sensitive to speaker information: (1) the relatively high C_{llr} values (local maxima) point to the least speaker-sensitive sub-bands (*filled in pink*) ranging roughly between 300 Hz and 2200 Hz; (2) the relatively low C_{llr} values (local minima) indicate the most speaker-sensitive sub-bands (*filled in blue*) located around low-, mid- and high-frequency ranges.

Overall, speaker information appears to be more strongly encoded towards the *high-frequency* sub-bands, as depicted in the Average panel where the C_{llr} pattern drops to its minimum from about 2300 Hz. While the *low-frequency range* (roughly below 500 Hz) tends to be less sensitive than the high-frequency one, it does carry notable speaker information albeit with vowel-to-vowel variations in C_{llr} values. Previous studies [6,13] have highlighted the relevance of this low range for speaker classification, especially targeting the back vowels of Japanese. In a similar vein, [16] have reported that the range [100-300 Hz] carries significant speaker information based on Japanese sentences spoken at normal, slow and fast rate, and [17] have found the range [0-770 Hz] to be highly useful for idendifying speakers with various accents of British English.

Fig. 2 also shows how the speaker-sensitive sub-bands identified through FVC are positioned with respect to formant regions. While exact alignment may not be expected partly due to statistical uncertainty in formant measurement, the speaker-sensitive sub-bands tend to fall near or within the marked F1, F2 and F3 regions and, in most cases, extend into likely F4 regions. This trend is reassuring with regard to the formant-dependent properties of vowels but, more importantly, it underscores the advantage of using the BLCC method to focus on any sub-band regions of potential relevance to speaker-specific information.

4.2. Experiment 2

In addition to identifying the locations of speaker-sensitive sub-bands as per the above, it is of interest to gauge their relative importance and performance with respect to the full band.

Experiment 2 thus consisted of fusing LRs for combinations of 2 to 7 sub-bands, with a view to observing FVC performance with an increasing number of sub-bands, and determining the combinations that yield optimal results. The best C_{llr} values per vowel are plotted in Fig. 3 for single sub-bands at left and, then, for fused sub-bands. The major improvements in C_{llr} occur with 3 fused sub-bands corresponding to those highlighted (in blue) in Fig. 2, in the same order from low to high-frequency regions.

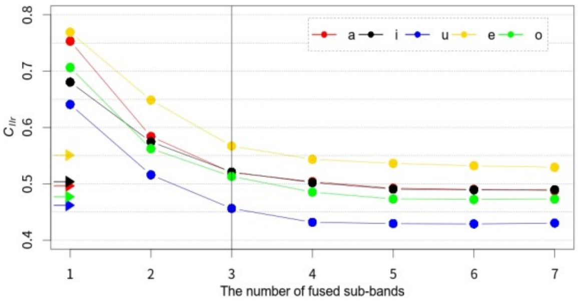


Figure 3: *Per-vowel Cllr values for the most speaker-sensitive sub-bands: single (1) and fused (2-7). Left arrows indicate Cllr values at full-band.*

The vertical line in Fig. 3 marks the point at which the inclusion of high-frequency sub-bands causes C_{llr} values to approach those obtained at full band, thus reinforcing our earlier observation regarding the significant role of those sub-bands. The speaker sensitivity of the high-frequency regions of our cepstrally-smoothed spectra agrees with previous acoustical studies of Japanese [13,18] and English [15,19] vowels. Shape differences in the lower part of the vocal tract (laryngeal tube and piriform fossa) are thought to cause the high-frequency range of speaker variation [20].

5. Concluding discussion and future work

This work has illustrated the efficacy of mel-frequency BLCCs and the usefulness of FVC in investigating sub-band regions of vowel spectra that are sensitive to speaker diffrences.

There is a clear consistency in the overall distributional pattern of speaker sensitivity and the locations of the most informative sub-bands (fused) with respect to relatively low C_{llr} values. For all vowels, two of these sub-bands respectively span the low- and the high-frequency ranges, while the other sub-band tends to be around the middle of the full range.

The relative performance of these sub-bands (quantified in experiment 2) sheds some light on the question raised in the introduction regarding MFCC spectra. Speaker information is clearly detectable in low- and mid-frequency ranges where spectral resolution is enhanced on the mel scale. More notably, our results indicate the effectiveness of MFCCs in capturing strong speaker differences in the vowels' high-frequency regions. Yet, this is where the frequency resolution of the mel scale is lower, and where spectral distinctiveness between speakers would expectedly be reduced. This apparent contradiction warrants further study, especially since the success of MFCCs in speech and speaker classification is commonly attributed to the very properties of the mel scale.

To that end, it may be beneficial to repeat the same FVC experiments using BLCCs on alternative frequency scales that differ in filter allocation: (1) BLCCs extracted from inverse-MFCCs [21], i.e., with the filters at high frequencies shifted to low frequencies, and vice versa; and (2) BLCCs extracted from linear-frequency CCs (LFCCs), i.e., with uniformly-spaced filters. In either case, it may be worthwhile to increase the number of filters (set to 25 in this work) and observe the impact of enhanced frequency resolution throughout the full range. Ultimately, some valuable insights may be gained from the patterns of speaker variances (within and between) in high-frequency sub-bands spaced on linear and mel scales.

The current study can conceivably be extended to non-vowel sounds [22] and female speech [23], which have been less studied in forensic context. The BLCC method should also be useful for locating the sub-bands that are more robust under adverse conditions, such as with mobile phones [24] and noisy environments, or for studying the impact of emotion-dependent sub-bands [25] on speaker variability.

Finally, it is hoped that the sub-band approach embedded in BLCCs will provide new perspectives in other areas of speech science and technology, where detailed investigation of spectral regions is required in an efficient and flexible manner. Possible areas of application extend from acoustic phonetics (e.g., the investigation of contrastive features [26,27]) and socio-phonetics (e.g., exploration of accent-specific sub-bands [28]) to spoofing detection [29,30], where synthetic traits may be notably encoded in different spectral regions.

6. Acknowledgements

The authors thank two anonymous reviewers for their valuable comments.

7. References

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366. 1980.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition from features to supervectors," *Speech Communication*, vol. 52, pp. 12-40. 2010.

[3] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Systems with Applications*, vol. 90, pp. 250–271. 2017.

[4] F. Clermont, "Linear transformation from full-band to sub-band cepstrum," in, *Proceedings of the 18th Australasian International Conference on Speech Science and Technology*, 2022, pp. 136-140.

[5] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, 1st ed. (Prentice Hall signal processing series). Englewood Cliffs, N.J.: Prentice Hall, 1993.

[6] S. Ishihara and F. Clermont, "The sub-band cepstrum as a tool for local spectral analysis in forensic voice comparison," in, *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, 2023, pp. 40-50.

[7] H. Makinae, T. Osanai, T. Kamada, and M. Tanimoto, "Construction and preliminary analysis of a large-scale bone-conducted speech database," *The Institute of Electronics, Information and Communication Engineers (IEICE) Technical Report*, vol. 107, no. 165, pp. 97-102, 2007.

[8] C. G. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, vol. 53, no. 1, pp. 109-122. 2004.

[9] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio," *Australian Journal of Forensic Sciences*, vol. 45, no. 2, pp. 173-197. 2013.

[10] Y. Kinoshita, T. Osanai, and F. Clermont, "Sub-band cepstral distance as an alternative to formants: Quantitative evidence from a forensic comparison experiment," *Journal of Phonetics*, vol. 94, no. 101177. 2022.

[11] G. E. Peterson, "The acoustics of speech–part II: Acoustical properties of speech waves." In L. E. Travis (ed.), *Handbook of Speech Pathology*, 1st ed., pp. 137-173. New York: Appleton-Century-Crofts, Inc., 1995.

[12] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," *Journal of the Acoustical Society of Japan (E)*, vol. 16, no. 5, pp. 283-289. 1995.

[13] M. Khodai-Joopari, F. Clermont and M. Barlow, "Speaker variability on a continuum of spectral sub-bands from 297-speakers' non-contemporaneous cepstra of Japanese vowels," in, *Proceedings of the 10th Australian International Conference on Speech Science and Technology*, 2004, pp. 504-509.

[14] R. Goto, K. Misawa and Y. Okada, "Analysis of individual characteristics in vowel spectral envelopes," In, *Proceedings of the International Multi-Conference of Engineers and Computer Scientists*, 2017, pp. 113-116.

[15] P. Mokhtari and F. Clermont, "A methodology for investigating vowel-speaker interactions in the acoustic-phonetic domain," in, *Proceedings of the 6th Australian International Conference on Speech Science and Technology*, 1996, pp. 127-132.

[16] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech Communication*, vol. 50, pp. 312–322. 2008.

[17] S. Safavi, A. Hanani, M. Russell, P. Jančovič and M. Carey, "Contrasting the effects of different frequency bands on speaker and accent identification," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 829-832. 2012.

[18] T. Kitamura and M. Akagi, "Relationship between physical characteristics and speaker individualities in speech spectral envelopes", in *Proceedings of 3rd Joint Meeting of the Acoustical Society of Japan and the Acoustical Society of Japan*, 1996, pp. 833-837.

[19] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 2, pp. 176-182. 1975.

[20] T. Kitamura, K. Honda and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects", in *Acoustical Science & Technology*, 2005, pp. 16-25.

[21] H. Lei and E. Lopez-Gonzalo, "Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition," in, *Proceedings of Inerspeech 2009*, 2009, pp. 2323-2326.

[22] P. Rose, "Likelihood ratio-based forensic semi-automatic speaker identification with alveolar fricative spectra in a real-world case," in, *Proceedings of the 18th Australasian International Conference on Speech Science and Technology*, 2022, pp. 6-10.

[23] P. Rose and E. Winter, "Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio analyses," in, *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, 2010, pp. 42-15.

[24] B. B. T. Nair, E. A. S. Alzqhoul and B. J. Guillemin, "Impact of the GSM and CDMA mobile phone networks on the strength of speech evidence in forensic voice comparison," *Journal of Forensic Research*, vol. 7, no. 324, pp. 1-9. 2016.

[25] M. H. Abed and D. Sztahó, "Effects of emotional speech on forensic voice comparison using deep speaker embeddings," in, *Proceedings of the XIX Hungarian Computational Linguistics Conference*, 2023, pp. 159-170.

[26] K. Iskarous, "The encoding of vowel features in mel-frequency cepstral coefficients." *Il parlato nel contesto naturale* [Speech in the Natural Context], 2018, pp. 9-18.

[27] M. Lambropoulos, F. Clermont and S. Ishihara, "The sub-band cepstrum as a tool for locating local spectral regions of phonetic sensitivity: A first attempt with multi-speaker vowel data," in, *Proceedings of Interspeech 2024*, 2024, pp. 1535-1539.

[28] L. M. Arslan and J. H. L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 28-40. 1997.

[29] M. H. Soni, T. B. Patel and H. A. Patil, "Novel subband autoencoder features for detection of spoofed speech," in, *Proceedings of Interspeech 2016*, 2016, pp. 1820-1824.

[30] B. Chettri, T. Kinnunen and E. Benetos, "Subband modeling for spoofing detection in automatic speaker verification," in, *Proceedings of Odyssey 2020: The Speaker and Language Recognition Workshop*, 2020, pp. 341-348.

Table of Contents for this manuscript

Preliminary Acoustic Analysis of Emotional Speech Production from New Zealand English Speakers with Parkinson's Disease

Itay Ben-Dom, Catherine I. Watson, Clare M. McCann

The University of Auckland

iben350@aucklanduni.ac.nz, c.watson@auckland.ac.nz , c.mccann@auckland.ac.nz

Abstract

This study explores the effects of Parkinson's disease (PD) on emotional speech production among native New Zealand English speakers, utilising a newly developed simulated emotional speech corpus. We engaged twelve participants diagnosed with PD to generate speech samples expressing five distinct emotions, yielding 1800 sentences. Our analysis focuses on key acoustic parameters, namely fundamental frequency and intensity, to evaluate their efficacy in depicting emotional states in PD-affected speech. Results indicate these parameters effectively differentiate between emotional expressions and correlate with emotional arousal levels. This provides insights into the complex interplay between physiological alterations due to PD and emotional vocal expression.

Index Terms: Emotional Speech Corpus, New Zealand English, Parkinson's disease, fundamental frequency, intensity.

1. Introduction

The human voice is a powerful conveyor of both biological and psychological information, capable of providing nuanced cues reflective of an individual's emotional state [1]. Acoustic and physiological cues in speech act as principal carriers of emotion, influencing sound production through alterations in muscle tension and respiration. These changes are key physiological indicators for affective vocalisation [2]. Emotional states profoundly impact voice production, necessitating modifications in respiration, phonation, and articulation, thereby requiring the quantification of these changes through various acoustic parameters to distinctly identify emotions [3].

Parkinson's disease significantly compromises the speech production mechanism through hypokinetic dysarthria, affecting the respiratory, phonatory, articulatory, and resonant systems. This leads to substantial challenges in production of emotional speech, where people with PD often exhibit reduced modulation of pitch and intensity, resulting in monotonous or hypophonic speech that negatively impacts their social communication and quality of life [4, 5].

F_0 and intensity are pivotal acoustic parameters for investigating vocal expressions [6]. F_0, the perceptual quality of pitch, is linked physiologically to vocal fold vibrations during speech production and varies across different emotional states. For example, higher F_0 values compared to the speakers' mean are associated with high-arousal emotions such as anger and fear, while lower values correspond to low-arousal emotions like sadness [7, 8, 9]. Intensity, or the loudness of the voice, also plays a critical role in emotional expression [4]. In PD, vocal intensity is often diminished due to weakened respiratory support and reduced control over vocal fold adduction, leading to softer speech that can further reduce speech intelligibility [10, 11]. These changes in intensity are critical for conveying emotional nuances and are influenced by muscular adjustments and respiratory effort, which are particularly compromised in PD. This results in a reduced dynamic range in vocal intensity, which can make emotional states more difficult to discern [12, 13, 14].

Despite advancements in understanding PD's effects on speech, there is a stark lack of research focusing on emotional speech production within this demographic, especially among native New Zealand English speakers. Existing speech corpora primarily feature other languages and are often limited to extracted acoustic features rather than comprehensive speech samples [15]. This gap underscores the necessity for a specialised corpus that addresses the unique needs of PD patients in New Zealand, facilitating detailed analysis of affective speech and its variations due to PD.

This study introduces the EMOPARKNZ corpus, developed to investigate the emotional speech production of New Zealand English speakers with PD. By analysing both fundamental frequency and intensity, this research aims to delineate the alterations in speech attributable to PD and provide insights into the complex interplay between physiological changes and vocal expression in affected individuals. This corpus represents a significant step toward understanding and improving communication aids for individuals with PD [15].

2. Methodology

2.1. Speakers

The corpus includes speech recordings from twelve individuals diagnosed with PD, balanced between six males and six females, all native speakers of New Zealand English. The age range for male participants is 56 to 74 years (mean 66.2 ± 6.2 years), and for female participants, 50 to 80 years (mean 64.3 ± 11.9 years). All participants were diagnosed by neurologists and recorded within 3 hours of medication intake to ensure they

Table 1. *Age, severity, and time after the PD diagnosis of male and female patients.*

M-PD			F-PD		
Age	**Severity**	**T**	**Age**	**Severity**	**T**
56	mild-mod	6	50	mild	4
62	mild-mod	8	53	mild	7
65	mild-mod	4	55	mild	6
67	mild-mod	20	74	mild-mod	3
73	mild	4	74	mild-mod	7
74	mild	14	80	mild-mod	10

M-PD: Male-PD, **F-PD**: Female-PD, **Severity**: mild/mild-moderate (mild-mod), and **T**: time post PD diagnosis (years).

were in the ON-state during speech sample collection, i.e. no more than 3 hours after taking medication [16]. Severity was determined by the researcher on the basis of overall impact of PD on daily functioning. Details regarding the age, severity, and years since onset diagnosis of each speaker are provided in Table 1.

2.2. Speech Materials

A speech corpus was developed to simulate strictly-guided emotions. Emotions were induced using a variation of the Stanislavski method [17], with specific s ituations a nd corresponding images presented to speakers to evoke and maintain the desired emotional state. This setup helped align the speakers' mental states with the targeted emotions throughout the recording session. Emotions were recorded in sequence from high to low arousal: *excited, angry, happy, neutral, sad*. To ensure consistency, the same leading phrases and images were provided to all speakers, who were instructed to maintain consistent emotional intensity. The corpus was designed with the JL corpus [18] in mind, a related NZ English speech data set developed for strictly-guided emotions. The sentences and prompts used in this study were adapted from those in the JL corpus. Fifteen different sentences, chosen for their neutrality across all primary emotions to prevent emotional bias, were included [19]. Short sentences of 4-7 syllables were selected to minimise potential emotional deviations within longer phrases. An example sentence is "Jack views an art piece." A complete list of all sentences used is shown in Table 2.

2.3. Recording process

Recordings for this study were approved by the University of Auckland Human Participants Ethics Committee (UAHPEC23603). The setup involved an omnidirectional condenser microphone (AKG, HC 577L) positioned 7 cm at a 45-degree angle to the right of the speaker's mouth. This microphone was connected to a Microsoft Surface Laptop 3 via a pre-amplifier (M-Audio MobilePre [MK II]), and audio capture was conducted at a sample rate of 44.1kHz with a 16-bit quantisation. The recording environment was prepared in a sound-treated room within the Speech Science laboratory at the University of Auckland, New Zealand. Participants were seated in a comfortable chair facing an iPad that displayed both the sentences and corresponding emotional images. This setup helped ensure participants were well-acquainted with the recording environment. Prior to recording, they were given five minutes to familiarise themselves with the settings and procedures, and they had the opportunity to ask any questions. The recording session commenced with a warm-up task where participants read the sentence "Tom beats that farmer" at a high intensity level. This initial task allowed for the adjustment of amplification gain levels to avoid clipping; these levels were set once and maintained throughout the session. Subsequently, the main recording task involved participants reading sentences displayed on the monitor, one at a time, each accompanied by an image corresponding to the intended emotion. The researcher monitored the recording in real-time to ensure technical quality. The sentence prompt was changed at the completion of each emotion expression. Each recording session lasted up to one hour, and the entire process was repeated to ensure two variations of each sentence for each emotion, resulting in a comprehensive corpus. In total, the study produced 1800 primary emotion sentences, encapsulating responses from 12 participants across 5 primary emotions, each repeated twice with 15 different sentences, all part of the EMOPARKNZ corpus.

Table 2. *Fifteen recorded sentences.*

No.	Sentence
1.	Tom beats that farmer.
2.	John laughs like your father.
3.	The seed is buried in deep.
4.	Taylor likes stewed Asian food.
5.	The lord swims in the sea.
6.	Jack views an art piece.
7.	Carl leaps into a jeep.
8.	Linda asks for more darts.
9.	Find your boot in this chute.
10.	Water harms the newborn boy.
11.	I have not seen my tooth.
12.	Work hard or you lose.
13.	Jim saw the port.
14.	They should start to talk.
15.	Sound the horn if you need more.

2.4. Data preparation

Each participant's acoustic recording was preserved in its entirety as a .wav file. Speech tasks were extracted from these files and stored individually to facilitate detailed analysis. The complete speech recordings were examined again for clippings and none were found, ensuring the quality of the data before further processing. The audio files were imported into Python (version 3.12.0) for initial processing, where leading and trailing silences were removed using the pydub package [20]. This refined acoustic signal was subsequently imported into MATLAB [21] for more advanced processing. Here, the fundamental frequency values of the speech samples were derived using the summation of residual harmonics method, implemented in the *covarep* package [22]. Additionally, intensity measurements (in dB) were extracted using the phonetic software Praat, employing its intensity function. Mean F_0 and intensity values were calculated for each sentence and aggregated for each speaker across all five primary emotions. After processing the signals and extracting the necessary acoustic features, a comprehensive statistical analysis was conducted. This analysis utilised the R language [23] and its environment for statistical computing and graphics (version 4.3.3), alongside the integrated environment R-Studio [24]. The data analysis aimed to statistically evaluate the relationships between extracted acoustic features and study variables, providing insights into the effects of Parkinson's disease on speech characteristics.

2.5. Statistical analysis

Our study employs a linear mixed-effects model, which is well-suited for data that include both fixed effects and random effects, allowing for the analysis of data with multiple sources of variability [25]. In this analysis, emotion (categorised as *excited, angry, happy, sad, neutral*) and gender (male/female) were treated as fixed effects. Speaker identity and the specific sentences they recited were considered as random effects, accommodating individual differences and sentence variability. The model formula used is represented in Equation 1 as follows:

$$\text{acoustic feature} \approx \text{emotion} * \text{gender} + (1|\text{subject}) + (1|\text{sentence}) \quad (1)$$

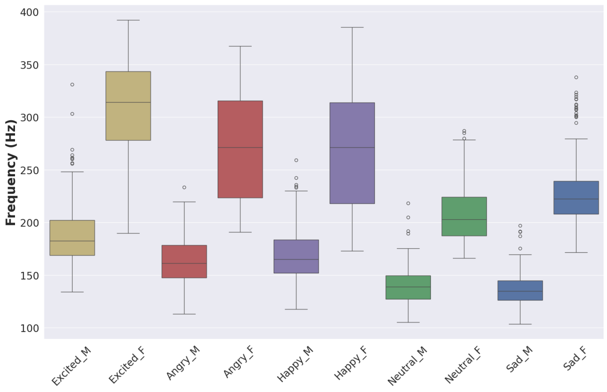


Figure 1: *Boxplot of F0 variation across five primary emotions for male (M) and female (F) speakers (n=12).*

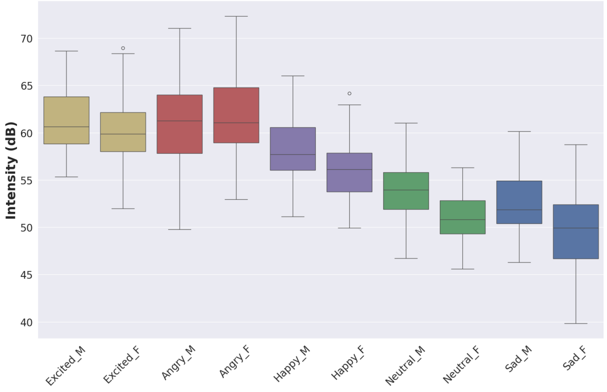


Figure 2: *Boxplot of intensity variation across five primary emotions for male (M) and female (F) speakers (n=12).*

The acoustic data was analysed using the linear mixed-effect model implemented in the R environment with the *lme4* package [26]. To refine the model and select the best fit, we used the step function from the *lmerTest* package [27], which simplifies the model selection process by comparing nested models. Subsequently, a likelihood ratio test was conducted to determine the statistical significance of the fixed effects, ensuring robustness in our findings. After determining the significant effects, post-hoc pairwise comparisons between levels of the fixed effects were performed. These comparisons were executed using the *emmeans* package [28], which adjusts p-values using the Tukey Honestly Significant Difference method to control for type I error across multiple tests. Effects were considered statistically significant if the p-values were less than 0.05, indicating robust differences in acoustic features across the groups defined by emotion and gender.

3. Results

3.1. Fundamental frequency

The analysis of F_0 across all twelve participants is depicted in Figure 1. Aggregated mean values and their standard deviations revealed the highest F_0 for the emotion *excited* (247.8 ± 65.2 Hz), followed in sequence by *happy* (219.8 ± 61.1 Hz), *angry* (216.1 ± 64.0 Hz), *sad* (182.8 ± 50.7 Hz), and *neutral* (173.7 ± 39.5 Hz). Statistical analyses utilising a linear mixed-effects model indicated significant interactions between most emotion pairs, except between *angry* and *happy*, detailed in Table 3. The linear mixed-effects model identified a statistically significant two-way interaction between emotion and gender, with a Chi-square value of $\chi^2_{(4)} = 1195$ and a p-value less than 0.0001. As expected, analysis revealed that females consistently exhibited higher F_0 values compared to males across all emotions studied. Specifically, for the emotion *excited*, females demonstrated an average F_0 of 306.9 Hz (±48.8 Hz), substantially higher than males who averaged 188.8 Hz (±31.4 Hz). Similarly, for *angry*, female speakers had an average F_0 of 270.6 Hz (±50.2 Hz), compared to 161.6 Hz (±24.2 Hz) in males. The significance continued with *happy* (females: 269.7±54.8 Hz, males: 169.9±26.2 Hz), *neutral* (females: 208.0±28.4 Hz, males: 139.4±19.0 Hz), and *sad* (females: 229.4±34.2 Hz, males: 136.1±16.8 Hz).

Table 3. *Statistical analysis comparison of emotion pairs.*

Emotion Pairs	Fundamental Frequency		Intensity	
	t-ratio	p-value	t-ratio	p-value
angry - excited	-15.413	**<.0001**	3.242	**0.0106**
angry - happy	-1.808	0.3692	22.139	**<.0001**
angry - neutral	20.604	**<.0001**	47.601	**<.0001**
angry - sad	16.195	**<.0001**	55.078	**<.0001**
excited - happy	13.605	**<.0001**	18.897	**<.0001**
excited - neutral	36.017	**<.0001**	44.360	**<.0001**
excited - sad	31.607	**<.0001**	51.836	**<.0001**
happy - neutral	22.412	**<.0001**	25.462	**<.0001**
happy - sad	18.002	**<.0001**	32.939	**<.0001**
neutral - sad	-4.410	**0.0001**	7.477	**<.0001**

3.2. Intensity

Corresponding analyses of acoustic intensity are illustrated in Figure 2. The mean intensity values, aggregated across all speakers, demonstrated the highest levels for the emotion *angry* (61.2 ± 3.6 dB), with descending values noted for *excited* (60.6 ± 2.7 dB), *happy* (57.1 ± 2.7 dB), *neutral* (52.4 ± 2.5 dB), and *sad* (51.0 ± 3.2 dB). The linear mixed-effects model revealed significant interactions across all pairs of emotions, detailed in Table 3. The linear mixed-effects model demonstrated a statistically significant two-way interaction between emotion and gender, with a Chi-square value of $\chi^2_{(4)} = 1195$ and a p-value less than 0.0001. Analyses revealed higher intensity values in males compared to females for all emotions, with the exception of *angry*. Specifically, for *excited*, intensity levels were higher in males (61.2 ± 3.4 dB) compared to females (60.0 ± 3.0 dB). Conversely, for *angry*, females exhibited a slightly higher intensity (61.6 ± 4.0 dB) than males (60.8 ± 4.4 dB). This trend persisted across other emotions, with males showing higher intensities for *happy* (58.3 ± 3.3 dB for males versus 56.0 ± 2.9 dB for females), *neutral* (53.9 ± 2.7 dB for males versus 50.9 ± 2.5 dB for females), and *sad* (52.6 ± 2.9 dB for males versus 49.4 ± 3.6 dB for females).

4. Discussion

Our acoustic analysis of the EMOPARKNZ corpus has demonstrated a systematic influence of arousal levels on fundamental frequency, confirming that higher arousal emotions such as *anger*, *excitement*, and *happiness* are characterised by elevated pitch compared to *neutral* speech. The observed trend shows fundamental frequency serves as an arousal-differentiating feature, as previously reported [18]. Interestingly, *sad* speech exhibited a higher average pitch than *neutral*, which deviates from common findings and suggests variations in the definition of neutral across different studies [29]. This may indicate that the emotional state labelled as "neutral" in this context could have involved subtle emotional undertones not typically accounted for in standard classifications.

Our findings reveal that high arousal emotions consistently show elevated mean pitch, likely due to increased tension in the cricothyroid muscle and heightened sub-glottal pressure [30, 31], which are necessary for producing the vocal intensity associated with these emotions. In contrast, lower arousal emotions like sadness are associated with a reduced mean pitch, which can be attributed to stiffer vocal folds and limited vibratory capacity [29]. These physiological responses highlight the impact of PD on the vocal apparatus's ability to modulate speech according to emotional demands. The study also uncovered significant gender differences in vocal expression, with females consistently displaying higher mean pitch across all emotions. This gender disparity may be influenced by anatomical differences in the laryngeal size, which are exacerbated by PD, affecting how emotions are vocally expressed between males and females [32].

The majority of studies report elevated mean fundamental frequency in individuals with PD compared to healthy controls [13, 32]. While data from a control group is not currently available, it is possible to compare the F_0 values for *neutral* speech with the mean F_0 values reported in the literature. The mean F_0 values for male and female healthy speakers are approximately 120 Hz and 180 Hz, respectively [33]. The results for EMOPARKNZ indicate mean F_0 values for male and female speakers to be 140 Hz and 208 Hz, respectively. These findings corroborate previous reports of elevated mean fundamental frequency in individuals with PD. This increase in fundamental frequency can be attributed to laryngeal muscular impairment due to rigidity [34].

In terms of intensity, our analysis indicated that emotions typically associated with high arousal were distinguished by their higher mean intensity levels, aligning with the increased energy demands these emotions impose on speakers. The observed trend shows that like fundamental frequency, intensity serves as an arousal-differentiating feature. The significant variability in intensity for emotions like anger and sadness suggests that individual differences in the interpretation of these emotions can substantially affect their acoustic expression [35, 36]. For example, anger can vary from cold to hot anger, impacting intensity levels differently, while the unexpected variability in sadness among professional actors points to possible methodological nuances in how these emotions were elicited during recordings [37]. Furthermore, we found notable differences in mean intensity between male and female speakers, reinforcing the need for gender-specific approaches in clinical settings [10, 11]. These findings are crucial for developing more effective therapeutic strategies that accommodate the unique speech and voice characteristics of men and women with PD.

While intensity demonstrated greater discriminatory power compared to fundamental frequency, it presents potential issues due to the variable placement of the microphone relative to the speaker's mouth, especially across different datasets. In contrast, fundamental frequency can always be derived from speech samples, making it a robust and popular acoustic descriptor. This consistency in measurement makes fundamental frequency particularly advantageous for machine learning training, as it provides a reliable and stable feature across diverse datasets.

The limitations of the EMOPARKNZ corpus arise from the scarcity of participants and the inherent challenges of the data collection process. PD is a movement disorder, which makes recruiting participants difficult, especially due to the associated anxiety and stress. Anxiety, a common non-motor symptom of PD, often leads to the avoidance of everyday social situations out of fear of embarrassment caused by PD symptoms [38]. PD symptoms tend to be exacerbated by stress when required to perform new tasks, further discouraging participants from volunteering for studies. With regards to the data collection process, the EMOPARKNZ corpus includes speech samples elicited through induced emotions. This approach allowed participants to mentally prepare themselves to produce the targeted emotions and to repeat sentences if necessary. While this method facilitates data collection, it may not accurately reflect the subtleties of emotional expression in natural conversation. Capturing natural emotional speech is challenging, as such expressions are often more subtle and harder to detect. During the interviews that preceded each recording session, participants expressed frustration about their emotions not being perceived accurately in conversations, particularly with unfamiliar people. When asked to demonstrate the difficulties they face in conveying emotions, several participants mentioned that when producing *angry* speech, they need to "talk louder in their head" to make their voice sound louder. This suggests that while fundamental frequency and intensity have been shown to be strong acoustic descriptors of emotion, it will be interesting to examine their effectiveness in natural emotional speech.

Overall, the results from this study not only enhance our understanding of how PD influences emotional speech but also underscore the importance of considering both fundamental frequency and intensity as key parameters in assessing and treating speech disorders associated with neurological conditions. The insights gained here pave the way for future research to explore additional acoustic features and apply these findings in clinical practice, potentially through advanced speech therapy techniques and the development of supportive communication technologies tailored to the needs of people with PD.

5. Conclusion and Future Directions

This study introduced the EMOPARKNZ corpus, a new strictly-guided simulated emotional speech corpus in New Zealand English from speakers with PD, highlighting the critical roles of F_0 and intensity in distinguishing emotional expressions in PD-affected speech. Acoustic analysis demonstrated the variability of prosody parameters across emotions, affirming their efficacy in capturing emotional content despite vocal impairments associated with PD. Notably, participants retained some ability to modulate their voices according to emotional context, with significant differences observed between genders. Future research will build upon these findings by conducting a comprehensive acoustic analysis incorporating temporal, spectral features, and glottal-based features. The EMOPARKNZ corpus will also be used to train neural network models for speech emotion recognition tasks, contributing to the development of user-facing healthcare robots. Upon completion of the author's doctoral thesis, the EMOPARKNZ corpus will be made available, contributing further to the scientific community's efforts to support people with PD in maintaining effective communication and improving their social interactions.

6. References

[1] A. Karpf, *The human voice: The story of a remarkable talent*. London: Bloomsbury, 2006.

[2] A. Kappas, U. Hess, and K. R. Scherer, "Voice and emotion," in *Fundamentals of nonverbal behavior*, R. S. Feldman and B. Rime, Eds. Cambridge University Press, 1991, ch. 6, p. 200–238.

[3] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1, pp. 227–256, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639302000845

[4] J. Möbes, G. Joppich, F. Stiebritz, R. Dengler, and C. Schröder, "Emotional speech in parkinson's disease," *Movement Disorders*, vol. 23, no. 6, pp. 824–829, 2008.

[5] M. Pell, H. Cheang, and C. Leonard, "The impact of parkinson's disease on vocal-prosodic communication from the perspective of listeners," *Brain and language*, vol. 97, pp. 123–34, 2006.

[6] I. R. Murray and J. L. Arnott, "Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech," *Computer Speech & Language*, vol. 22, no. 2, pp. 107–129, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230807000393

[7] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[8] C. Breitenstein, D. V. Lancker, and I. Daum, "The contribution of speech rate and pitch variation to the perception of vocal emotions in a german and an american sample," *Cognition and Emotion*, vol. 15, no. 1, pp. 57–79, 2001.

[9] D. Guo, H. Yu, A. Hu, and Y. Ding, "Statistical analysis of acoustic characteristics of tibetan lhasa dialect speech emotion," *SHS Web of Conferences*, vol. 25, p. 01017, 2016.

[10] C. Dromey, L. O. Ramig, and A. B. Johnson, "Phonatory and articulatory changes associated with increased vocal intensity in parkinson disease: A case study," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 4, pp. 751–764, 1995.

[11] Cynthia M. Fox and Lorraine Olson Ramig, "Vocal sound pressure level and self-perception of speech and voice in men and women with idiopathic parkinson disease," *American Journal of Speech-Language Pathology*, vol. 6, no. 2, pp. 85–94, 1997.

[12] R. J. Holmes, J. M. Oates, D. J. Phyland, and A. J. Hughes, "Voice characteristics in the progression of parkinson's disease," *International Journal of Language & Communication Disorders*, vol. 35, no. 3, pp. 407–418, 2000.

[13] C. Dromey, "Spectral measures and perceptual ratings of hypokinetic dysarthria," *Journal of Medical Speech-Language Pathology*, vol. 11, pp. 85–94, 06 2003.

[14] J. E. Huber and M. Darling, "Effect of parkinson's disease on the production of structured and unstructured speaking tasks: Respiratory physiologic and linguistic considerations," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 1, pp. 33–46, 2011.

[15] C. Williams and K. Stevens, "Emotions and speech: some acoustical correlates." *The Journal of the Acoustical Society of America*, vol. 52 4, pp. 1238–50, 1972.

[16] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease," in *International Conference on Language Resources and Evaluation*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:1735319

[17] J. Benedetti, *Stanislavski: An Introduction, Revised and Updated*, 2nd ed. Routledge, 2004. [Online]. Available: https://doi.org/10.4324/9780203998182

[18] J. James, L. Tian, and C. Watson, "An open source emotional speech corpus for human robot interaction applications," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*, B. Yegnanarayana, Ed. Hyderabad: ISCA, September 2018, pp. 2768–2772.

[19] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond emotion archetypes: Databases for emotion modelling using neural networks," *Neural Networks*, vol. 18, no. 4, pp. 371–388, 2005, emotion and Brain. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608005000353

[20] J. Robert, M. Webbie *et al.*, "Pydub," http://pydub.com/, 2018.

[21] T. M. Inc., "Matlab version: 9.13.0 (r2022b)," Natick, Massachusetts, United States, 2022. [Online]. Available: https://www.mathworks.com

[22] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[24] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2020. [Online]. Available: http://www.rstudio.com/

[25] B. Winter, "Linear models and linear mixed effects models in r with linguistic applications," 2013.

[26] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, p. 1–48, 2015. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v067i01

[27] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmertest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, p. 1–26, 2017. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v082i13

[28] R. V. Lenth, P.-C. Buerkner, M. Herve, J. Love, F. Miguez, H. Riebl, and H. Singmann, "emmeans: Estimated marginal means, aka least-squares means," https://CRAN.R-project.org/package=emmeans, 2022.

[29] J. James, "Modeling prosodic features for empathetic speech of a healthcare robot," Ph.D. dissertation, University of Auckland, 2021.

[30] A.-M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen, "Physical variations related to stress and emotional state: a preliminary study," *Journal of Phonetics*, vol. 24, no. 3, pp. 313–335, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447096900170

[31] Y. Li, J. Li, and M. Akagi, "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. 908–916, 2018.

[32] S. Skodda, H. Rinsche, and U. Schlegel, "Progression of dysprosody in parkinson's disease over time — a longitudinal study," *Movement Disorders*, vol. 24, no. 5, pp. 716–722, 2009.

[33] J. Clark, C. Yallop, and J. Fletcher, *An introduction to phonetics and phonology*, 3rd ed. Malden, Mass: Blackwell Pub, 2007, vol. Series: Blackwell textbooks in linguistics.

[34] I. Midi, M. Dogan, M. Koseoglu, G. Can, M. A. Sehitoglu, and D. I. Gunal, "Voice abnormalities and their relation with motor dysfunction in parkinson's disease," *Acta Neurologica Scandinavica*, vol. 117, no. 1, pp. 26–34, 2008.

[35] T. Johnstone and K. Scherer, "The effects of emotions on voice quality," *Proceedings of the XIVth International Congress of Phonetic Sciences*, p. 5, 1999.

[36] K. Scherer, "Vocal affect expression: a review and a model for future research," *Psychological bulletin*, vol. 99 2, pp. 143–65, 1986.

[37] J. Wilting, E. J. Krahmer, and M. Swerts, "Real vs. acted emotional speech," in *Interspeech*, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:14024275

[38] B. K. Gultekin, B. Ozdilek, and E. E. Bestepe, "Social phobia in parkinson's disease: Prevalence and risk factors," *Neuropsychiatric Disease and Treatment*, vol. 10, pp. 829–834, May 21 2014.

Exploring Fundamental Frequency Characteristics of Australian Adolescents with and without Depression in the Future Proofing Study Corpus

Tünde Szalay[1], Brian Stasak[1,2], Kate Maston[2], Aliza Werner-Seidler[2], Mark Larsen[2,3]

[1]School of Elect. Eng. & Telecomm., [2]Black Dog Institute, [3]Centre for Big Data Research in Health
University of New South Wales

t.szalay@unsw.edu.au

Abstract

Lower mean and reduced fundamental frequency ($F0$) variation are speech-related markers of depression in adults. Using the Future Proofing Corpus, $F0$ differences in contours, variation, and creakiness were examined in read aloud sentences produced by 14 Australian adolescent female speakers with and without depression. Contrary to known $F0$ markers of depression, the current study found no evidence for speakers with depression producing lower or less varied $F0$ than speakers without depression, potentially due to other factors (e.g., age, puberty) masking depression-related $F0$ differences. However, speakers with depression were significantly more likely to produce creaky voice than their peers without depression.

Index Terms: Australian English, corpus phonetics, creaky voice, $F0$, mental health, paralinguistics

1. Introduction

Depression has a prevalence rate of 15% among Australian adolescents [1]. The Future Proofing Study (FPS) examines factors associated with mental health conditions, including depression, during adolescence using big data [1]. The FPS collected demographic and mental health information using surveys, as well as audio recordings to capture speech markers of mental health conditions [1, 2]. The FPS corpus contains speech data suitable for acoustic analysis of changes in speech associated with depression.

Changes in speech, such as lower speech rate, increased hesitation markers, and lower and less varied intonation, are frequent speech markers of depression [3, 4]. In adults, clinical literature describes the voice of speakers with depression as "low", "monotonous", and "toneless" due to listeners auditory-impressionistic perception of patients with depression as having lower and less varied pitch [4]. In line with the perceived low pitch of speakers with depression, overall fundamental frequency ($F0$) was quantitatively found to reduce with depression severity in adults [5, 6]. Similarly to adults, adolescent male speakers with depression produced lower mean $F0$ compared to adolescents without depression when discussing a series of topics, such as event planning or conflict resolution [7]. In contrast, adolescent female speakers with depression produced higher $F0$ values compared to their peers without depression, but only when discussing event planning [7].

Supporting perceptual clinical observations on depressed voice being "monotonous", speakers with depression show relatively narrower range, less variation, and slower rate of changes in $F0$ in acoustic studies [5, 8]. Speakers with depression produce less $F0$ variation based on context and topic compared speakers without depression [5, 8]. For example, speakers with depression have less $F0$ variability differentiating happy, sad, or angry speech [5]. Speakers with more severe depressive symptoms produce less $F0$ variation to distinguish clear speech from conversational speech compared to speakers with less severe symptoms [8]. Reduced $F0$ variation can be explained by increased muscle tension in the larynx caused by psychomotor retardation, i.e., the slowing of thought and reduction of physical movements accompanying depression [4]. However, increasing $F0$ range and velocity with increasing depression severity has also been found by acoustic studies [9].

Inconsistent results are attributed to $F0$ being affected by multiple factors besides depression, such as feeling agitated or anxious [7, 10]. For example, patients produced lower $F0$ after depression treatment compared to pre-treatment due to feeling more relaxed post-treatment [10]. Despite the inconsistent acoustic results, perceived pitch may be considered during clinical assessments, and $F0$ measurements have contributed to the development of speech-based automatic depression screening algorithms achieving approximately 70% accuracy [3, 11].

The majority of studies showing $F0$ differences associated with depression were carried out on American English adult [4] and adolescent speech [7, 11]. Despite potential accent differences in $F0$, automatic depression detection algorithms can successfully classify speech as depressed or non-depressed using $F0$ characteristics of Australian English (AusE) speech [12, 13].

The FPS corpus provides self-reported symptoms of depression combined with mobile phone speech data suitable for developing automatic depression detection algorithms for a population that may be difficult to access: young adolescents reporting symptoms of depression. To improve validity and interpretability of such algorithms, we examined $F0$ patterns of Australian adolescents with and without depression in the FPS corpus. We tested whether $F0$ markers of depression present in adults are also present in the young adolescents in the FPS corpus, hypothesising that AusE-speaking adolescents reporting symptoms of depression would produce lower $F0$ and reduced $F0$ variation.

2. Methods

2.1. Corpus description

The FPS corpus contains survey data from 6,388 participants and 14,498 sound files elicited from 1,062 year 8 students recruited from 134 secondary schools across Australia and collected across multiple timepoints [2]. Participants undertook the data collection activities on their own time using their smartphones with no obligation to complete the speech recordings, resulting in the different number of participants submitting survey- and audio data.

Survey data included demographic information, such as participants' age, linguistic- and cultural background; health information concerning disabilities, drug-consumption and COVID-19; and a series of mental health symptoms. Symptoms of depression were measured using the Patient Health Questionnaire for Adolescents (PHQ-A) [14]. PHQ-A ranges from 0 to 27, with higher scores indicating more severe depression; PHQ-A $\geq$ 15 was used as the threshold for caseness [1].

Two protocols were used for speech elicitation. The first protocol included the diadochokinetic /pataka/ task, the Harvard sentences [15], the Rainbow passage [16] and was used from August 2019 to April 2021. The second protocol included a sustained /a:/; diadochokinetic /pataka/; a prosodic sentence set designed to elicit different intonation contours associated with describing movement; an affective sentence set with positive, negative, and neutral valence designed to elicit emotions; open-ended question; categorisation; and a creative task asking participants to list homonyms. The second protocol was used from April 2021 onwards, until the end of data collection.

Students submitted survey and speech data in a self-timed, unsupervised manner at a baseline timepoint and at eight follow-up points [2]. Speech and survey data were submitted to two different servers and linked via timestamps. The data collection method allowed for delayed submission of speech samples relative to the survey data. That is, depression severity may have been different at the time of audio recording than what was reported in the survey.

2.2. Corpus data extraction

To reduce changes in $F0$ associated with participants going through puberty during real-time data collection, only sound files submitted within 31 days of baseline survey data collection were considered in the analysis. To reduce $F0$ variation between speakers due to non-depression-related factors, only Australian-born cisgender speakers who reported speaking English at home and reported no disabilities were included.

To select the most suitable speech tasks, we considered constraints of phonetic and clinical validity. Only read speech tasks were considered as they had a uniform format, length, and contents. For clinical validity, we only considered sentence- or passage reading tasks. To increase the likelihood of capturing $F0$ differences, we selected the tasks designed to elicit large $F0$ range and variation. The prosodic sentence set was designed to elicit $F0$ variation via the juxtaposition of contrasting motion- and speed-related words (e.g., *The hare quickly went down the hill* vs. *The bear slowly went up the hill*), consisting of 24 paired sentences. Out of the 12 pairs, one pair (two sentences) were assigned randomly to each participant. The affective sentence set contained 10 positive, 10 negative, and 10 neutral sentences, designed to elicit a variety of $F0$ patterns due to $F0$ being a key marker of emotions [17]. Out of the 30 sentences, one positive, one negative, and one neutral sentences were assigned randomly to each participant. The prosodic and affective sentence tasks combined provide five sentences per participant.

Out of the 1,061 speakers submitting audio files, 781 submitted audio at baseline, 574 met speaker inclusion criteria, and 69 submitted the prosodic and affective sentence tasks. Of the 69 speakers, 15 (M = 5, F = 10) reached the threshold for caseness of depression. Male speakers were excluded due to the low number of male speakers with depression in the sample. Three female speakers with depression were excluded due to low audio quality, and seven were included. The seven included speakers with depression (F = 7) had a mean age of 13.46 years and reported diverse sexualities (lesbian/bi-/pansexual = 5, heterosexual = 2). As lower $F0$ is expected to mark both age and depression [18], the 7 adolescents with depression were matched with 7 female adolescents without depression according to age (mean = 13.34 years). As lesbian sexual orientation may be indexed in other languages by lowered and less varied $F0$ [19, 20], the markers of depression under investigation, the 7 adolescents without depression matched the group with depression according to sexuality (diverse = 5, heterosexual = 2).

In line with analysis of the survey responses of the baseline cohort (6,388 participants), speakers with depression were more likely to be female than male and more likely to identify with diverse sexualities than identify as heterosexual among participants submitting speech data [1].

2.3. *F*0 analysis

A total of 14 (speakers) × 5 (sentences) = 70 sentences were included in the analysis. $F0$ analysis followed the protocol recommended for semi-automatic $F0$ estimation [21]. Recorded sentences were force-aligned automatically using the Montreal Forced Aligner with a British pronunciation dictionary and a multi-accented acoustic model; manual correction was not completed to improve replicability [22, 23]. $F0$ values were estimated automatically and overlaid on spectrograms for visual inspection at every 37.5 ms between 20 Hz to account for creaky voice and 600 Hz to account for adolescent female speakers having high $F0$ [21, 24]. $F0$ estimates were time-aligned with the automatic phonetic segmentation; $F0$ values estimated during non-sonorant segments were excluded [21]. $F0$ variation was captured as the absolute value of $F0$ difference between $F0$ measured at the end of a sonorant interval and $F0$ measured at the start of the following sonorant interval.

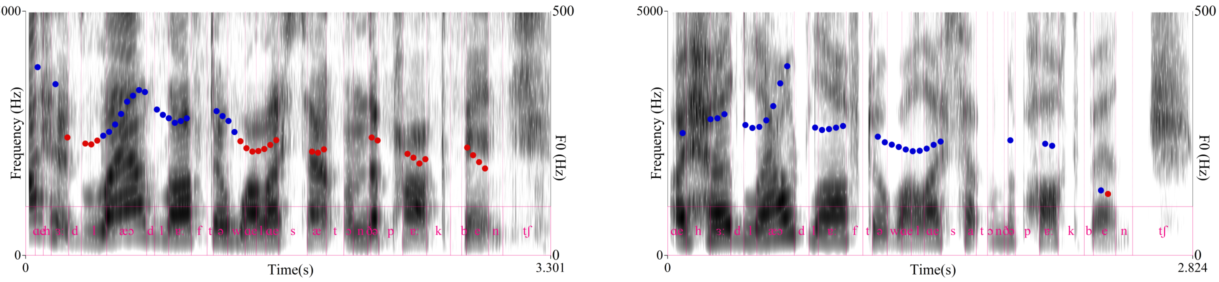
2.3.1. *Automatic creak detection*

As young AusE speakers are likely to produce creaky voice, automatic creak detection was applied to the data by visualising each speaker's $F0$ distribution using histograms [21, 25, 26]. Creaky voice was identified in bimodal distributions,by the presence of a second peak with a lower mode, separated from the higher peak of modal voice by an anti-mode [26, 21].

The number of modes was determined using the excess mass test in the *antimode* library of R and visual inspection [27, 28, 29]. When both excess mass test and visual inspection indicated bimodal $F0$ distribution, the speaker was considered to have produced creaky and modal voice. When the excess mass test indicated bimodal $F0$ distribution, but visual inspection suggested that the $F0$ distribution was right-skewed rather than bimodal, the speaker was considered to have produced modal voice. When neither the excess mass test nor the visual inspection indicated bimodal distribution, the speaker was considered to have produced modal voice. $F0$ values produced below the anti-mode in bimodal distributions were rated as creaky (Figure 1a). For all other speakers, $F0$ values below 132 Hz threshold were rated as creaky (Figure 1b) [26].

Average $F0$ of modal voice in the current FPS data produced by 12-14-year-old female speakers fell between that of 11-year-old prepubescent female speakers and 18-29-year-old adult female speakers (11-year-old = 248 Hz; 12-14-year-old = 237 Hz, 18-29-year-old = 206 Hz) [18, 26].

Table of Contents for this manuscript



(a) *Predominantly creaky voice produced by a speaker with depression.* 60% of the F0 estimates identified as creaky and 40% as modal.

(b) *Predominantly modal voice produced by a speaker without depression.* 2.9% of F0 estimates identified creaky and 97.1% as modal.

Figure 1: *The sentence* I heard loud laughter while I sat on the park bench *produced by two speakers, overlaid with automatic segmentation (pink) and* F0 *contour (creaky: red, modal: blue).*

2.4. Statistical analysis

As automatic creak detection indicated a potential difference in creak between speakers with and without depression, $F0$ produced during modal and creaky voice were included in the statistical analysis. In addition, differences in the proportion of creaky voice between speakers with and without depression were examined statistically, despite not being hypothesised.

To test $F0$ differences between speakers with and without depression, one linear mixed model and two generalised linear mixed models were constructed. The response variable $F0$ was modelled using a linear mixed model with Gaussian family. The response variable $F0$ Variation was modelled using a generalised linear mixed model with Gamma family, as $F0$ Variation had a right-skewed distribution. The presence of creak was modelled using generalised mixed model with binomial family with the response variable Creaky (modal $F0$ coded as 0, creaky $F0$ coded as 1). In all models, the independent variable was Depressed (categorical, comparing “With Depression” to the baseline “No Depression”) with a by-participant random effect to account for intraspeaker variation, such as age, puberty, and sexual orientation. All models were constructed using the *lme4* library in R; p-values were calculated with the *lmerTest* library using Satterthwaite’s degrees of freedom method [30, 31].

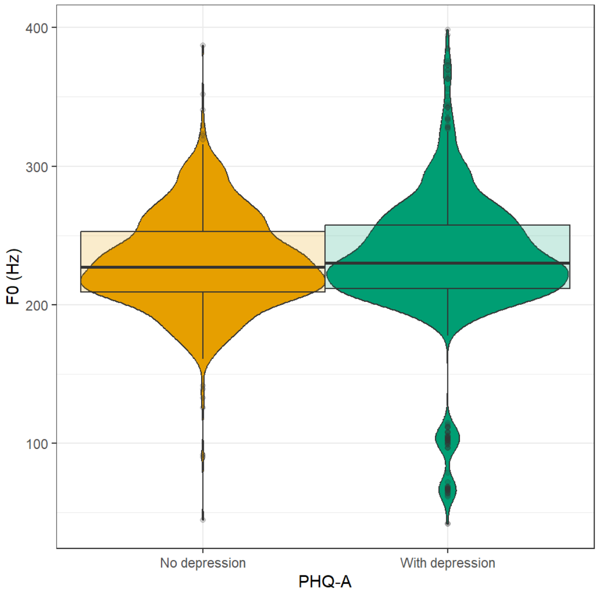


Figure 2: F0 *produced by speakers with and without depression; random effect of speaker not visualised.*

3. Results

Speakers with and without depression did not show a significant difference in $F0$ and $F0$ variation ($F0$: $\beta = -1.45$ Hz, $t_{12.71} = -0.11, p = 0.911$, Fig. 2; $F0$ variation: $\beta = -0.01$ Gamma-transformed Hz, $t_{0.01} = -1.31, p = 0.19$, Fig. 3).

Speakers with depression produced a significantly higher proportion of creaky voice compared to speakers without depression ($\beta = 3.92\%, z_{1.59} = 2.46, p = 0.0138$, Fig. 4). Four out of seven speakers with depression produced bimodal $F0$ distribution and three produced unimodal distribution. Out of the three speakers with depression producing unimodal distribution, one produced $F0$ values below the 132 Hz threshold rated as creaky, and two did not produce $F0$ values rated as creaky at all. All seven speakers without depression produced unimodal $F0$ distribution, with two producing values below the threshold of creaky voice, and five producing $F0$ rated as modal.

4. Discussion

The study aimed to test if $F0$ markers of depression were present in the FPS corpus of Australian adolescents. Adolescents reporting symptoms of depression were hypothesised to produce lower and less varied $F0$. Contrary to the hypotheses, we did not find differences in overall $F0$ or in $F0$ variation between speakers with and without depression. That is, lowered and less varied $F0$ symptoms were either not present in the current sample of the FPS corpus or were too small to discover.

Changes in $F0$ accompanying depression are expected have small effect size, as shown by the inconsistent $F0$ differences in the literature [4]. In our study, small effects of depression could have been masked by other factors, such as more robust puberty related differences. Although groups with and without depression were balanced for age, adolescents may not go through puberty at the same time, thus, one group may have more speakers who have reached puberty than the other. The potential effects of sexual orientation may also mask $F0$ symptoms of depression if adolescents reporting diverse sexual orientation use lower and less varied $F0$ [19, 20]. However, effects of sexual orientation cannot be reliably tested in the current sample.

Exploratory analysis revealed that adolescent female speak-

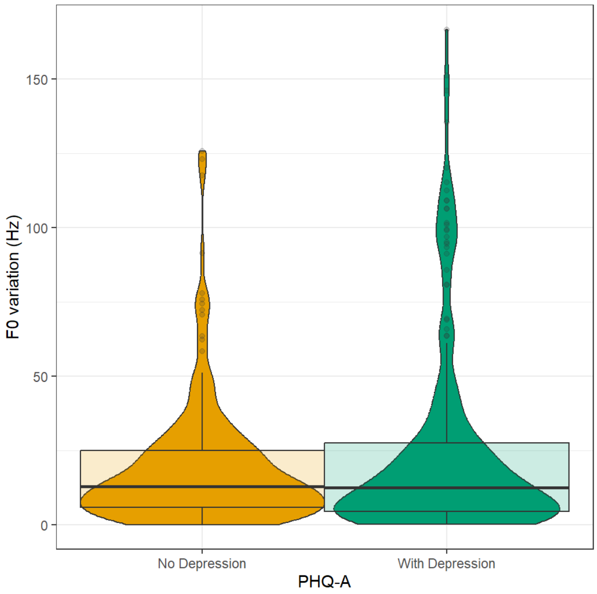


Figure 3: *F0 variation produced by speakers with and without depression; random effect of speaker not visualised.*

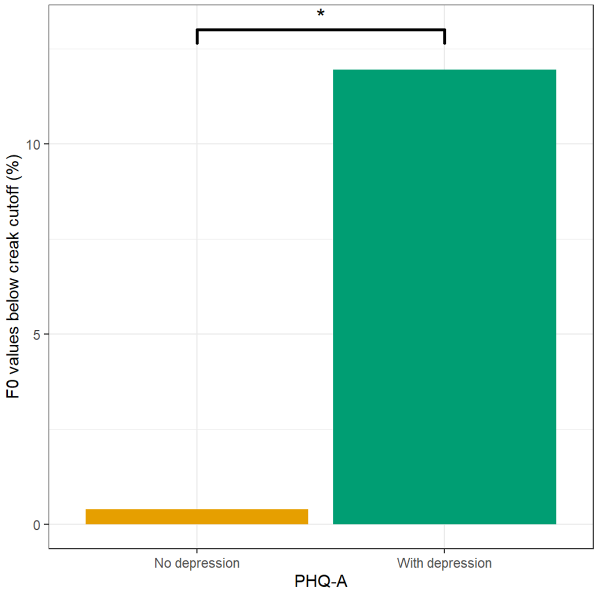


Figure 4: *Percentage of* F0 *estimates below the cutoff-point of creak relative to the total number of* F0 *estimates per group; random effect of speaker was not visualised.*

ers with depression produced a higher proportion of creaky voice compared to the speakers not reaching the threshold of depression. That is, depression may be marked by intra-speaker variation as speakers lower their *F*0 relative to themselves using creak rather than relative to their peers without depression. Low open quotient, another characteristic of creaky voice, has also been identified as a speech marker of depression [32, 33]. In contrast, when voice quality was captured using harmonics-to-noise ratio and jitter, increased depression severity was marked by aspiration and breathiness, indicating breathy, rather than creaky voice [9]. Results on voice quality as a marker of depression are difficult to compare between studies due to lack of standardisation in extraction techniques [4]. Future research may examine depression and creak using a variety of acoustic characteristics of creaky voice [33].

As *F*0 produced during creaky voice is lower than *F*0 produced during modal voice [33], increased creak among speakers with depression could have resulted in the expected overall lower *F*0. Despite the increase in creak, *F*0 lowering was not observed, potentially due to the relatively low proportion of creaky to modal voice (Figure 4). Low prevalence of creak is attributed to speaker- and task effects. The current sample only contained female speakers who are less likely to produce creaky voice than male speakers [34]. Creak is often used a socio-indexical marker [34]; as speech was elicited via read sentences, the speakers might not have produced such markers.

The lack of difference in overall *F*0 and *F*0 variation is inconsistent with *F*0 differences contributing to the successful classification of speech as belonging to speakers with and without depression [12, 13]. However, increased presence of creaky voice may also contribute to automatic depression detection via the acoustic correlates of creak, such as low and irregular *F*0, high noise, or high harmonics-to-noise ratio [33]. Future research developing automatic depression screening algorithms using the FPS corpus is required to identify the most suitable speech features for automatic depression classification in the FPS corpus. As automatic classification algorithms consider multiple features rather than just *F*0, *F*0 and creak may be suitable to classify speakers as depressed or not depressed when combined with other speech factors.

The main limitation of the current study is the small sample size. While the overall corpus is large, containing 14,498 sound files from 1,062 participants, only 70 sentences from 14 participants were analysed in the current study. The number of suitable speakers reduced substantially due to extracting only the affective sentences and prosodic sentences tasks submitted at baseline. As baseline data were collected from August 2019 to March 2022, and the second protocol containing affective and prosodic sentences was introduced in April 2021, fewer participants submitted the affective- and prosodic sentences tasks compared to speech tasks from the first protocol, such as The Rainbow passage. Although design and selection of the affective and prosodic sentences tasks were motivated by capturing *F*0 variation, a key speech symptom of depression, the low number of speakers submitting these tasks is likely to have contributed to the null results. Future research may explore the use of other speech tasks, for instance The Rainbow passage, to examine the links between *F*0, *F*0 variation, creak, and depression in the FPS corpus.

5. Conclusion

We analysed *F*0 differences between AusE-speaking adolescent female speakers with and without symptoms of depression using the Future Proofing Study corpus. The speakers included in this study did not exhibit differences in overall *F*0 or *F*0 variation. It is possible that the relatively small effects of depression on *F*0 were masked by other, more robust factors, such as puberty. Speakers with depression; however, produced a higher proportion of creaky voice. Future work will expand the current analysis to test differences in the use of creaky and modal voice between speakers with and without depression using more data from the Future Proofing Study Corpus.

6. Acknowledgements

The Future Proofing Study was approved by the University of New South Wales (NSW) Human Research Ethics Committee (HC180836), the State Education Research Applications Process for the NSW Department of Education (SERAP2019201), and relevant Catholic Schools Dioceses across Australia. This project was supported by the NHMRC GNT1197074 and GNT1152952 grants to Aliza Werner-Seidler and NHMRC GNT1138405 and GNT1155614 grants to Helen Christensen. The authors thank Helen Christensen for her work, and Hannah White and Timothy Shea for their comments.

7. References

[1] A. Werner-Seidler, K. Maston, A. L. Calear, P. J. Batterham, M. E. Larsen, M. Torok, B. O'Dea, K. Huckvale, J. R. Beames, L. Brown *et al.*, "The future proofing study: Design, methods and baseline characteristics of a prospective cohort study of the mental health of Australian adolescents," *International Journal of Methods in Psychiatric Research*, vol. 32, no. 3, p. e1954, 2022.

[2] A. Werner-Seidler, K. Huckvale, M. E. Larsen, A. L. Calear, K. Maston, L. Johnston, M. Torok, B. O'Dea, P. J. Batterham, S. Schweizer *et al.*, "A trial protocol for the effectiveness of digital interventions for preventing depression in adolescents: The future proofing study," *Trials*, vol. 21, no. 1, pp. 1–21, 2020.

[3] K. R. Scherer, "Vocal assessment of affective disorders," in *Depression and expressive behavior*, J. D. Maser, Ed. New York: Routledge, 1987, pp. 57–82.

[4] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[5] Z. Breznitz, "Verbal indicators of depression," *The Journal of general psychology*, vol. 119, no. 4, pp. 351–363, 1992.

[6] J. Wang, L. Zhang, T. Liu, W. Pan, B. Hu, and T. Zhu, "Acoustic differences between healthy and depressed people: a cross-situation study," *BMC psychiatry*, vol. 19, pp. 1–12, 2019.

[7] A. Y. Hussenbocus, M. Lech, and N. B. Allen, "Statistical differences in speech acoustics of major depressed and non-depressed adolescents," in *2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS)*. IEEE, 2015, pp. 1–7.

[8] H. Yi, R. Smiljanic, and B. Chandrasekaran, "The effect of talker and listener depressive symptoms on speech intelligibility," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 12, pp. 4269–4281, 2019.

[9] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity." in *Interspeech*, vol. 2, 2012, pp. 1059–1062.

[10] F. Tolkmitt, H. Helfrich, R. Standke, and K. R. Scherer, "Vocal indicators of psychiatric treatment effects in depressives and schizophrenics," *Journal of communication disorders*, vol. 15, no. 3, pp. 209–222, 1982.

[11] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2010.

[12] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. Parker, M. Breakspear *et al.*, "Characterising depressed speech for classification," 2013.

[13] B. Stasak, J. Epps, and R. Goecke, "Automatic depression classification based on affective read sentences: Opportunities for text-dependent analysis," *Speech Communication*, vol. 115, pp. 1–14, 2019.

[14] J. G. Johnson, E. S. Harris, R. L. Spitzer, and J. B. Williams, "The patient health questionnaire for adolescents: validation of an instrument for the assessment of mental disorders among adolescent primary care patients," *Journal of Adolescent Health*, vol. 30, no. 3, pp. 196–204, 2002.

[15] E. H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[16] G. Fairbanks, "Voice and articulation drillbook," *(No Title)*, 1960.

[17] T. Johnstone and K. R. Scherer, "Vocal communication of emotion," *Handbook of emotions*, vol. 2, pp. 220–235, 2000.

[18] P. A. Busby and G. L. Plant, "Formant frequency values of vowels produced by preadolescent boys and girls," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 2603–2606, 1995.

[19] B. Moonwomon-Baird, "Toward a study of lesbian speech," *Queerly phrased: Language, gender, and sexuality*, pp. 202–213, 1997.

[20] J. Van Borsel, J. Vandaele, and P. Corthals, "Pitch and pitch variation in lesbian women," *Journal of Voice*, vol. 27, no. 5, pp. 656–e13, 2013.

[21] H. White, J. Penney, A. Gibson, A. Szakay, and F. Cox, "Evaluating automatic creaky voice detection methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 3, pp. 1476–1486, 2022.

[22] M. McAuliffe and M. Sonderegger, "English (UK) MFA dictionary v3.0.0," https://mfa-models.readthedocs.io/pronunciationdictionary/English/English(UK)MFAdictionaryv3_0_0.html, Tech. Rep., Feb 2024.

[23] ——, "English MFA acoustic model v3.0.0," https://mfa-models.readthedocs.io/acoustic/English/EnglishMFAacousticmodelv3_0_0.html, Tech. Rep., Feb 2024.

[24] P. Boersma and D. Weenink, "Praat 6.4.01,," 2023. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[25] J. Penney, F. Cox, and A. Szakay, "Glottalisation, coda voicing, and phrase position in Australian English," *The Journal of the Acoustical Society of America*, vol. 148, no. 5, pp. 3232–3245, 2020.

[26] Y. Leung, J. Oates, V. Papp, and S.-P. Chan, "Speaking fundamental frequencies of adult speakers of Australian English and effects of sex, age, and geographical location," *Journal of Voice*, vol. 36, no. 3, pp. 434–e1, 2022.

[27] J. Ameijeiras-Alonso, R. M. Crujeiras, and A. Rodriguez-Casal, "Multimode: An R package for mode assessment," *arXiv preprint arXiv:1803.00472*, 2018.

[28] J. Ameijeiras-Alonso, R. M. Crujeiras, and A. Rodríguez-Casal, "Mode testing, critical bandwidth and excess mass," *Test*, vol. 28, pp. 900–919, 2019.

[29] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org/

[30] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, p. 1–48, 2015.

[31] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, p. 1–26, 2017.

[32] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd." in *Interspeech*, 2013, pp. 847–851.

[33] P. A. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice." in *ICPhS*, vol. 1, 2015, pp. 2–7.

[34] H. White, J. Penney, A. Gibson, A. Szakay, and F. Cox, "Creak prevalence and prosodic context in Australian English," in *Annual Conference of the International Speech Communication Association (24th: 2023)*. International Speech Communication Association (ISCA), 2023, pp. 112–116.

Impact of Speech Emotions on Vowel Space

Daniel Cho, Catherine Watson, Jesin James

Department of Electrical, Computer, and Software Engineering,
The University of Auckland, New Zealand
dcho981@aucklanduni.ac.nz, c.watson@auckland.ac.nz, jesin.james@auckland.ac.nz

Abstract

This study explores how speech emotions affect vowel spaces, focusing on valence and arousal dimensions across two speech corpora: German Emotional Speech Corpus (EMO-DB) and New Zealand English Emotional Speech Corpus (JLCorpus). Formant extraction on the primary and secondary emotions from JLCorpus gives the F1 and F2 frequencies for vowels. Results show high arousal emotions correlate with higher mean F1 values, while low arousal emotions correspond to lower mean F1 values. Positive valence emotions tend to have higher mean F2 values. Statistical analyses confirm significant emotional influences on vowel formants, supporting existing research and extending insights into secondary emotions.

Index Terms: Formants, emotional speech, vowel space, primary emotions, secondary emotions

1. Introduction

People can recognise emotional states through facial expressions, body movements, gestures and speech [1]. The focus of this study is emotion expression via speech. A listener's ability to recognise affective states just by listening to vocal cues is based on the fact that human speech contains prosody, spectral and other suprasegmental features which convey emotional information [2]. The importance of pitch and loudness as acoustic features that differentiate emotions is well established [3, 4, 5]. Here, we contribute to the studies which look at the impact that emotions have on articulation by focusing on the resonances of the vocal tract, known as formant frequencies.

As the resonances of the vocal tract are not fixed, any change in the position of the tongue and the jaw will change the resonance of the vocal tract. Specifically, the first and second formants can be related to the jaw opening and tongue movement during articulation, respectively. As the tongue moves forward and backward, it changes the length and shape of the vocal tract. Generally, higher formant frequencies (e.g., F2 and F3) correspond to the position of the tongue body (front to back). A more open jaw creates a longer vocal tract, which tends to lower all formant frequencies (F1, F2, F3, etc.). This is because longer vocal tracts resonate at lower frequencies. Hence, visualising the vowels in a 2D space where first and second formants represent the y and x axes respectively (called the vowel space) is used as an approach to relate the formant values to the jaw opening and tongue movement during articulation. This makes the vowel space, and the formants a powerful tool to study factors that impact articulation. Among the other articulation strategies, lip rounding tends to lower F1 frequency of vowels. This is because when the lips are rounded, they create a constriction in the vocal tract that effectively increases its length acoustically. A longer vocal tract resonates at lower frequencies, so F1 decreases when the lips are rounded [6]. Formants, influenced by speech production physiology, can contribute to analyses of mental health [7, 8, 9], speech disorders [10], and speaker characteristics [11], but cannot be viewed as standalone indicators.

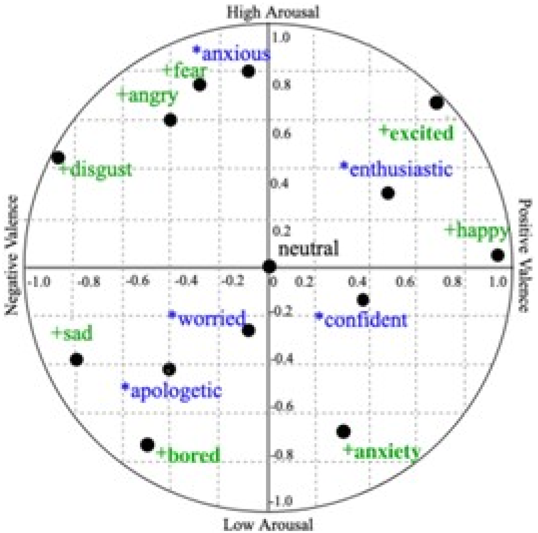


Figure 1: *Valence-arousal plane showing the emotions in the EMO-DB and JLCorpus, primary emotions are labelled with '+' and secondary emotions are labelled with '*' [12]*

Emotions also impact articulation, and hence analysing the formant frequencies can provide insights into how the vocal tract is modified during emotion production. The authors in [1] were able to extract more valuable emotional content by segmenting vowels from utterances and analysing the formant frequencies of the vowels. Other studies [13, 14, 15] also report an improvement in emotion classification accuracy when formants are used as features. While these studies provide valuable insights into the impact of emotions on formants, detailed analysis on the effect of emotions on each of the vowels can provide insights into articulatory changes during emotion production. Hence, the research questions answered in this paper are:

1. How do formant frequencies of vowels change during the expression of emotions?
2. How do the articulatory modifications of the vocal tract necessary for emotive speech vary based on the levels of arousal and valence associated with emotions?

In this paper, the first two formants (F1 and F2) will be focused on as these two formants can be used to differentiate vowels, thereby teasing out the impact that emotions have on vowel production. Using a dimensional emotion model - Russel's circumplex model of emotions [16], a visual representation of the emotions on a valence arousal (V-A) two-dimensional plot is shown in Figure 1 [17]. The valence dimension indicates pleasantness (E.g., happy having positive valence, sad having negative valence) and the arousal dimension indicates the level

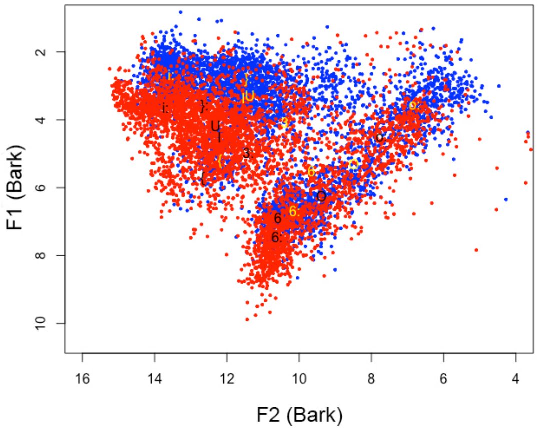


Figure 2: *Untransformed male (blue) and female (red) vowels from JLCorpus and the centroids, male (yellow), female (black)*

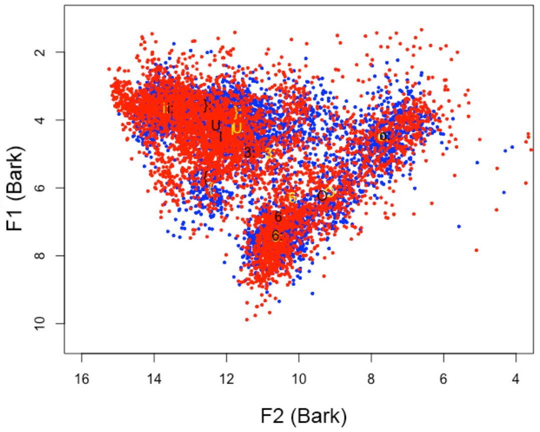


Figure 3: *Transformed male (blue) and female (red) vowels from JLCorpus and the centroids, male (yellow), female (black)*

of reaction to a stimulus (E.g., excited having high arousal, sad having low arousal. In the V-A dimension as seen in Figure 1, an emotion such as excited has positive valence, high arousal, while an emotion such as angry has negative valence, high arousal. This V-A two-dimensional representation will be used to interpret the impact that the valence and arousal levels of the emotions have on the formant frequencies, and thereby the articulation.

2. Methodology

2.1. Emotional Speech Corpora

Two speech databases are used to investigate the effects of speech emotion on formats - German Berlin Emotional Speech Database EMO-DB [18] and New Zealand English JLCorpus [12]. EMO-DB consists of seven emotions (anger, boredom, anxiety, happiness, sadness, disgust and neutral) conveyed by ten professional actors (five male and five female) through ten German utterances. In the JLCorpus, 10 emotions are divided into two sections: primary and secondary. The primary emotions are angry, excited, neutral, happy, and sad. The secondary emotions are anxious, apologetic, confident, enthusiastic and worried. EMO-DB and JLCorpus have a variety of high and low arousal emotions as well as positive and negative valence emotions which are labelled in Figure 1. To understand the placement of emotions along the arousal dimension, arrange them vertically from anxious (high arousal) at the top to bored (low arousal) at the bottom. For the valence dimension, arrange emotions horizontally from happy (positive valence) on the right to sad (negative valence) on the left. Although secondary emotions are nuanced in contrast to primary emotions [12], the significance of secondary emotions in social interactions is undeniably substantial and therefore must be studied [19, 17]. The emotions within the JLCorpus are well spread amongst the valence-arousal plan allowing for an extended scope compared to EMO-DB. EMO-DB has 12 German monophthongs distributed within a total of 535 utterances used for formant analysis in this paper. Whereas the JLCorpus has 11 New Zealand English monophthongs distributed within a total of 2400 utterances that are used for formant analysis. Sampling frequency rates are 16 kHz for EMO-DB and 44.1 kHz for JLCorpus.

2.2. Segmentation and Formant Extraction

EMO-DB speech recordings are labelled at the word and phonetic levels using forced alignment, enabled by the Munich Automatic Web Segmentation System, webMAUS [20] using the German language setting. The labelled database was then converted to an EMU-formatted database [21] for exact phonetic querying and segmentation. After segmentation, first two formants F1 and F2 are extracted for the vowels /aː/, /eː/, /ɛː/, /iː/, /ɪ/, /oː/, /uː/, /yː/ and /ʏ/ using the signal processing function forest within the R package wrassp [22]. The default forest parameters are used for the male speakers and the nominal F1 was set to 600 Hz for the female speakers. The formant frequencies for each vowel instance are time normalized and extracted from temporal midpoints.

JLCorpus recordings were segmented, and formants were extracted using the same method as EMO-DB, although the New Zealand English setting was used for webMAUS. Formant extraction was conducted for the vowels /iː/, /ɪ/, /ɛ/, /æ/, /ɒː/, /ɔ/, /oː/, /uː/, /ɜː/. There was no hand correction of phonetic boundaries and formant tracks for EMO-DB and JLCorpus data.

2.3. Data Transformation

Due to differences in vocal tract length between the speakers identifying as male and female, it can be challenging to empirically compare the vowel spaces of the speakers [23]. Therefore, for statistical analysis, a linear transformation is performed to make formant frequency observations comparable. This transformation allows for a conversion of male speaker values closer to female speaker values or vice versa; the choice is made by the researchers. The transformation method was first proposed in [24] and has also been used in other studies e.g. [25]. Both EMO-DB and JLCorpus have equal number of female and male speakers in their databases. A male-to-female transformation was conducted as reported in [25].

Figure 2 shows the male (blue) and female (red) vowel spaces respectively for all the vowel data in the JLCorpus before the transformation was performed. Figure 3 shows the transformed male vowel space and the female vowel space. The anchoring vowels for the JLCorpus were /iː/ /aː/ and /oː/. A similar transformation process was conducted on the male data of the EMO-DB with the anchoring vowels for this corpora being /iː/, /aː/ and /uː/. All formant plots in this paper show the female data and the transformed male data on the same plots.

3. Results

The F1 and F2 values are analysed separately for each vowel and presented in a centroid vowel space in bark frequency scale.

Table of Contents for this manuscript

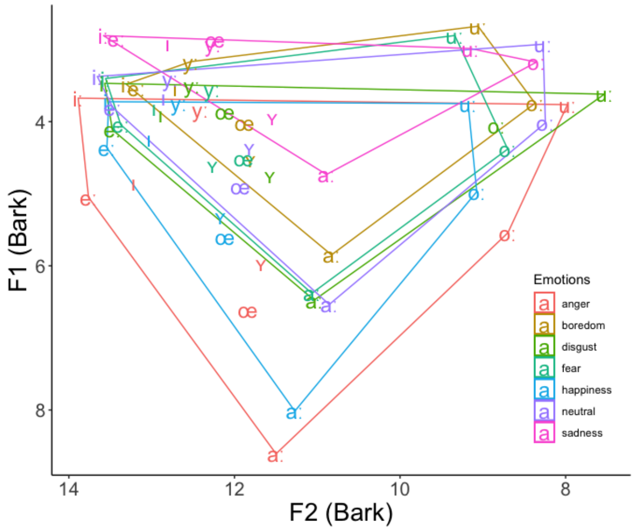


Figure 4: *Vowel space of all German speakers after transformation in different emotions from EMO-DB*

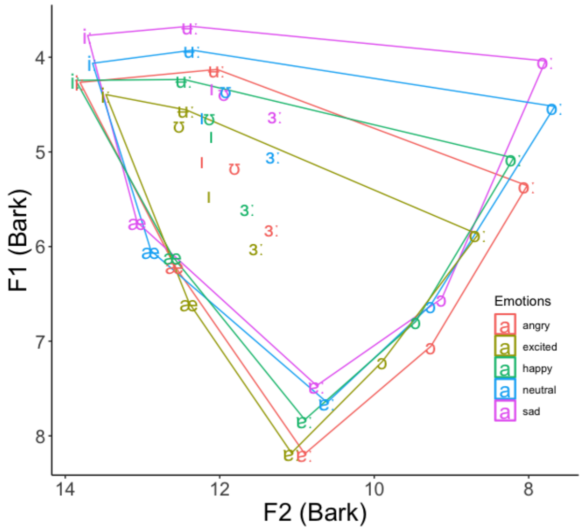


Figure 5: *Vowel space of the primary emotions after transformation of all New Zealand English speakers from JLCorpus*

3.1. Analysis of Emotions in EMO-DB Database

3.1.1. Vowel Space Analysis

Figure 4 depicts the German vowel space across different emotions. We will compare the vowel space with the relative valence-arousal values of the emotions shown in Figure 1. It is evident that the F1 axis effectively distinguishes between high and low arousal emotions, ranging from anger (high arousal) to sadness (low arousal). Emotions characterized by high arousal such as anger, disgust, fear, and happiness exhibit higher F1 values, resulting in an expanded vowel space. Conversely, emotions with low arousal like boredom and sadness display lower mean F1 values, leading to a more condensed vowel space. These findings align with previous studies on formant analysis of EMO-DB [14], which also noted a strong correlation between mean F1 values and speakers' emotional arousal.

3.1.2. Statistical Analysis

Statistical analysis of individual vowels revealed significant effects when comparing models that included both formant type and emotions against a null model with fixed effects of formant type alone. This comparison aimed to ascertain whether there were statistically significant differences in vowels across emotions compared to neutral. Notably, vowels /aː/, /eː/, /iː/, /ɪ/, /oː/, /ʊ/, /yː/, /ʏ/ showed significant emotion-related impacts, evidenced by p-values below 0.05. Subsequently, pairwise comparisons using t-tests (Table 1) were conducted to pinpoint where these significant differences occurred among the fixed effects. In this analysis, neutral served as the reference.

Specifically, emotions such as anger, sadness, and happiness exhibited statistically significant differences in both F1 and F2 formant frequencies compared to neutral. This underscores the influence of these emotions have on vowel production across the dataset.

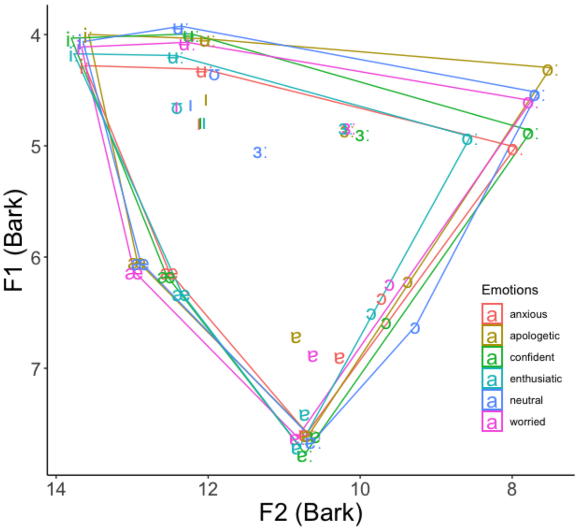


Figure 6: *Vowel space of the secondary emotions after transformation of all New Zealand English speakers from JLCorpus*

Table 1: *T Test of significant vowels impacted by emotion relative to neutral from EMO-DB*

Emotion	Vowel	F1/F2	Value	t-value	p-value
Anger	/iː/	F1	0.3	2.0	4e-2
Anger	/oː/	F1	1.5	7.5	3e-10
Anger	/oː/	F2	-1.1	-2.8	5e-3
Anger	/aː/	F1	2.01	4.5	3e-5
Anger	/aː/	F2	-1.4	-3.2	1e-3
Sadness	/iː/	F1	-0.6	-4.9	6e-6
Sadness	/iː/	F2	0.6	2.7	6e-3
Sadness	/oː/	F1	-0.8	-4.1	1e-4
Sadness	/oː/	F2	0.9	2.4	1e-2
Sadness	/aː/	F1	-1.9	-4.1	1e-4
Sadness	/aː/	F2	1.9	4.5	7e-6
Happiness	/iː/	F1	0.3	2.3	2e-2
Happiness	/iː/	F2	-0.5	-2.2	2e-2
Happiness	/oː/	F1	0.9	4.1	1e-4
Happiness	/aː/	F1	1.5	3.2	2e-3

3.2. Vowel Space of Emotions in JLCorpus Database

Figure 5 shows New Zealand English vowel space of the primary emotions from the JLCorpus. The relative location of each vowel within that space agrees with the general F1-F2 distribution of the New Zealand English vowels [26]. A substantial contrast between the vowel spaces of the primary emotions is seen which can be attributed to the different valence and arousal levels of each emotion. Higher aroused emotions such as angry, excited and happy have a similar vowel spread with a corresponding higher mean F1. At the same time, lower arousal emotions such as sad and neutral can be seen with a vowel space with an overall lower mean F1. Positive valence emotions happy and excited have a moderately higher mean F2, which can also be seen from the vowel centroid (can be quite clearly seen for the /oː/ vowel) shifted to the left more than other emotions.

The resulting New Zealand English vowel space of the secondary emotions in Figure 6 shows a similar trend regarding mean F1-F2 values relating to valence and arousal of emotions. The vowel space of the secondary emotions is not as spread out, therefore, it is more difficult to distinguish clear boundaries between each emotion's vowel space. Regardless, higher arousal emotions - notably anxious and enthusiastic have a higher mean F1 overall whereas, low arousal emotions namely apologetic and worried have an apparent lower mean F1. It is also evident that the positive valence emotion enthusiastic has a higher mean F2 than other negative valence emotions. Formant positioning is also significantly affected by the valence dimension as well as vowel position.

3.2.1. Statistical Analysis

Statistical analysis of the JLCorpus was performed in the same method as EMO-DB, where a fixed effects formant type model was compared with a model based on formant and emotions. Results indicate significant emotional impact on vowels /ɒː/, /iː/, /oː/ and /uː/ with p-values less than 0.05. Afterwards, t-tests are conducted for these vowels with neutral used as the emotion reference level. The results showed that the F1 values for angry, excited, happy, sad, anxious, and enthusiastic emotions are each (statistically) significantly different from those for neutral. Significant differences in F2 were found for angry, excited, sad, and anxious compared to neutral.

3.3. Discussion

[14] performed formant analysis on EMO-DB and identified a direct correlation between the F1/F2 positions at the vowel level and the level of the speaker's emotional arousal. The study was a gender-dependent evaluation of the complete list of German phonemes. In contrast to Vlasenko et al's research [14], the formant analysis in our study took a phonetic approach, emphasizing the acoustic characteristics of vowel production. Our approach involved visualizing vowel space with F1 on the y-axis and F2 on the x-axis to highlight the phonetic relationship between vowel articulation and acoustic properties. Also, our analysis was solely emotion focused, therefore, a data transformation of the formants was conducted due to the impact of the gender differences in vocal tract length on the formants.

The phonetic approach allows linking the vowel space to possible articulation strategies. For example, for the the vowel /ɒː/ from the JLCorpus, there is a clear difference between sad and excited in F1, suggesting that the jaw is more open for excited than sad. In contrast, there is hardly any difference in F1 of vowel /ɒː/ for the secondary emotions, suggesting jaw opening is not contrasting those emotions.

The vowel spaces pf EmoDB and JLCorpus datasets were obtained to be ordered based on the arousal levels of emotions, as shown in Figures 4, 5, and 6. While some effects of valence levels were observed, they did not reach statistical significance.

High arousal emotion angry can be seen with a larger vowel space and higher centre of gravity (centre of vowel space) in comparison to low arousal emotion sadness when the red centroid is compared to the magenta centroid in Figure 4. The overall vowel space and centre of gravity of low arousal emotions can be seen lifted (lower F1) compared to high arousal emotions and neutral.

In the German EmoDB dataset, all corner vowels (/i/, /a/, and /u/) demonstrate sensitivity to arousal levels. In New Zealand English, it was found that the back vowel /o/ is particularly effective in distinguishing between emotions, as depicted in Figures 5 and 6. Although differences were noted for other corner vowels like /i/ and /ɒ/, these distinctions were less pronounced compared to /o/. This suggests that the tongue position required to produce the back vowel is a crucial articulatory strategy for conveying varied emotional states. Additionally, the presence of lip rounding may serve as an articulatory strategy for low-arousal emotions (such as sadness), given that lip rounding typically lowers the F1 value [27]. It is important to note that the first formant of the /o/ vowel, due to its proximity to the fundamental frequency, often faces challenges in accurate formant estimation. Therefore, significant deviations in the /o/ vowel cannot be conclusively verified without manual verification of formant estimates.

The vowels /iː/, /eː/, /aː/, /oː/, /ʏ/, /yː/ and /ɪ/ from EMO-DB were found to be better at conveying emotional differences than others. Five of the seven vowels listed are long vowels, therefore, vowel length plays an important role in distinguishing between the different emotions. The NZE vowels /iː/ , /ɪ/ , /ɒː/ , /oː/ and /uː/ from the JLCorpus conveyed emotional differences better than other vowels. The vowels listed were all long vowels with the exception of /ɪ/, further verifying the findings with EmoDB that vowel length is an articulation strategy to differentiate emotions.

The addition of statistical analysis was necessary to make the results more robust. Due to formant analysis on emotions not being vastly popular, secondary emotion's impact on the vowel space has rarely been studied. Although secondary emotions are more subtle and nuanced than secondary emotions [17], the JLCorpus emotional speech database provided and introduced secondary emotions which were analysed in this paper. Results showed that the vowel space of subtle secondary emotions such as anxious and enthusiastic still displayed significant differences compared to the neutral vowel space.

4. Conclusion

Many acoustic features have been used for emotion analysis. Among the most important of these are features that describe the pattern of resonances within the vocal tract, also known as formants. Formant position is a spectral property of the speech signal that reflects voice quality as well as linguistic vowel identity. Most past research carried out in the field has tended to focus on a smaller set of rather strong emotions, such as anger, joy, sadness, and fear. To bridge this research gap, secondary emotions from the JLCorpus were used in this study. This study explored the the first and second formants in emotional speech and linked it with articulation, thereby answering the first research question on how formants are impacted during emotion production. We also related the valence and arousal values of the emotions to the formant values via vowel spaces, thereby answering the second research question linking the valence-arousal dimension to emotion expression.

In the future, research relating higher and lower mean F1 values to the vocal tract resonance such as past research generating vocal tract shapes from formant frequencies [28] would allow insight into how different emotions impact the jaw and tongue movements. While formant analysis can aid in classifying emotions, it should be combined with other methods due to the lack of a direct relationship between formants and emotions.

5. References

[1] M. Goudbeek, J.-P. Goldman, and K. R. Scherer, "Emotion dimensions and formant position." in *Interspeech*, vol. 2009, 2009, pp. 1575–1578.

[2] M. Meyer, M. Keller, and N. Giroud, "Suprasegmental speech prosody and the human brain," *The Oxford handbook of voice perception*, pp. 143–165, 2018.

[3] J. James, L. Tian, and C. Watson, "An open source emotional speech corpus for human robot interaction applications," *Interspeech 2018*, 2018.

[4] H. Nordström, "Emotional communication in the human voice," Ph.D. dissertation, Department of Psychology, Stockholm University, 2019.

[5] P. Laukka, H. A. Elfenbein, N. S. Thingujam, T. Rockstuhl, F. K. Iraki, W. Chui, and J. Althoff, "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features." *Journal of personality and social psychology*, vol. 111, no. 5, p. 686, 2016.

[6] R. D. Kent, J. Dembowski, and N. J. Lass, "The acoustic chracteristics of american english," in *Principles of Experimental Phonetics*, N. J. Lass, Ed. St. Louis, Missouri: Mosby, 1996, pp. 185 – 225.

[7] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Comparing objective feature statistics of speech for classifying clinical depression," in *The 26th annual international conference of the IEEE engineering in medicine and biology society*, vol. 1. IEEE, 2004, pp. 17–20.

[8] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[9] S. Scherer, L.-P. Morency, J. Gratch, and J. Pestian, "Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4789–4793.

[10] W. R. Rodríguez and E. Lleida, "Formant estimation in children's speech and its application for a spanish speech therapy tool." in *SLaTE*, 2009, pp. 81–84.

[11] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[12] J. James, L. Tian, and C. Watson, "An open source emotional speech corpus for human robot interaction applications," *Interspeech 2018*, 2018.

[13] J. C. Kim and M. A. Clements, "Formant-based feature extraction for emotion classification from speech," in *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2015, pp. 477–481.

[14] B. Vlasenko, D. Philippou-Hübner, D. Prylipko, R. Böck, I. Siegert, and A. Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal emotions," in *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 2011, pp. 1–6.

[15] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," *Information Sciences*, vol. 563, pp. 309–325, 2021.

[16] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[17] J. James, B. Balamurali, C. I. Watson, and B. MacDonald, "Empathetic speech synthesis and testing for healthcare robots," *International Journal of Social Robotics*, vol. 13, no. 8, pp. 2119–2137, 2021.

[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[19] J. James, C. I. Watson, and B. MacDonald, "Artificial empathy in social robots: An analysis of emotions in speech," in *2018 27th IEEE International symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2018, pp. 632–637.

[20] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[21] S. Cassidy and J. Harrington, "Multi-level annotation in the emu speech database management system," *Speech communication*, vol. 33, no. 1-2, pp. 61–77, 2001.

[22] R. Winkelmann, J. Harrington, and K. Jänsch, "Emu-sdms: Advanced speech database management and analysis in r," *Computer Speech & Language*, vol. 45, pp. 392–410, 2017.

[23] J. Harrington, "Acoustic phonetics," *The handbook of phonetic sciences*, pp. 81–129, 2010.

[24] C. Watson, B. Ross, E. Ballard, H. Charters, R. Arnold, and M. Meyerhoff, "Preliminary investigation into sound change in auckland," in *Proceedings of the 17th Australasian International Conference on Speech Science and Technology*, 2018, pp. 17–20.

[25] B. Ross, "An acoustic analysis of new zealand english vowels in auckland," Ph.D. dissertation, Open Access Te Herenga Waka-Victoria University of Wellington, 2018.

[26] C. I. Watson and A. Marchi, "Resources created for building new zealand english voices," in *Proc. 15th Australas. Int. Conf. Speech Science and Technology*, 2014, pp. 92–95.

[27] N. Lass, *Principles of Experimental Phonetics*. Mosby, 1996. [Online]. Available: https://books.google.co.nz/books?id=jUliAAAAMAAJ

[28] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," *The Journal of the Acoustical Society of America*, vol. 64, no. 4, pp. 1027–1035, 1978.

Author Index